JYVÄSKYLÄ STUDIES IN COMPUTING
262

Mirka Saarela

Automatic Knowledge Discovery from Sparse and Large-Scale Educational Data

Case Finland

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 262

Mirka Saarela

Automatic Knowledge Discovery from Sparse and Large-Scale Educational Data

Case Finland

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 3
kesäkuun 16. päivänä 2017 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 3, on June 16, 2017 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Automatic Knowledge Discovery from Sparse and Large-Scale Educational Data

Case Finland

JYVÄSKYLÄ STUDIES IN COMPUTING 262

Mirka Saarela

Automatic Knowledge Discovery from Sparse and Large-Scale Educational Data

Case Finland

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Editors
Timo Männikkö
Faculty of Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7084-0

URN:ISBN:978-951-39-7084-0
ISBN 978-951-39-7084-0 (PDF)

ISBN 978-951-39-7083-3 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2017

ABSTRACT

Saarela, Mirka
Automatic Knowledge Discovery from Sparse and Large-Scale Educational Data. Case Finland.
Jyväskylä: University of Jyväskylä, 2017, 98 p.(+included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; **262**)
ISBN 978-951-39-7083-3 (nid.)
ISBN 978-951-39-7084-0 (PDF)
Finnish summary
Diss.

The Finnish educational system has received a lot of attention during the 21st century. Especially, the outstanding results in the first three cycles of the Programme for International Student Assessment (PISA) have made Finland's education system internationally famous, and its unique characteristics have been under active research by various, predominantly educational, scholars since then. However, despite the availability of real but often sparse big data sets that would allow more evidence-based decision making, existing research to date has mostly concentrated on using classical qualitative and (univariate) quantitative methods. This thesis discusses, in general terms, knowledge discovery from large and sparse educational data—particularly from PISA—through the utilization and further development of multivariate data mining techniques and, more specifically, the application of these methods in the context of the Finnish educational system. Therefore, its goals are twofold and interrelated: to advance knowledge discovery methods and algorithms for sparse educational data to gain more interpretable models and to utilize these approaches to learn from the data and improve understanding of educational phenomena. This article-style dissertation is composed of 10 publications. The first publication provides a general knowledge discovery framework for analyzing sparse educational data. The succeeding seven publications discuss and advance methods for the special characteristics and complexities of PISA data and their usage for the quantitative educational knowledge discovery process. The final two publications demonstrate how human advising and decision making in Finnish educational institutions and related to the management of a national educational system can be automated and improved by employing the introduced analysis framework and process. All this provides new insights about Finnish education, advances the overall automatic quantitative knowledge discovery process, increases institutional awareness, and could save costs on various levels of the whole educational system.

Keywords: PISA, Knowledge Discovery, Sparse Data, Educational Data Mining, Learning Analytics, Educational Data Science, Finland, Big Data

Author

Mirka Saarela
Department of Mathematical Information Technology
University of Jyväskylä
Finland
mirka.saarela@jyu.fi

Supervisors

Professor Dr. Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Professor Dr. Jouni Välijärvi
Finnish Institute for Educational Research
University of Jyväskylä
Finland

Reviewers

Professor Dr. Petri Nokelainen
Industrial and Information Management
Tampere University of Technology
Finland

Professor Dr. Agathe Merceron
Media Informatics Department
Beuth University of Applied Sciences
Germany

Opponent

Professor Dr. Rolf Vegar Olsen
Centre for Educational Measurement
University of Oslo
Norway

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Prof. Dr. Tommi Kärkkäinen and Prof. Dr. Jouni Välijärvi for excellently mentoring this work. I want to thank Tommi for our regular long meetings and inspiring discussions, his active guidance, invaluable knowledge, continuous support, and encouragement at all stages of my thesis. Most of all, I want to thank Tommi for believing in me, for pushing me to next levels, and his genuine enthusiasm for joint work. I want to thank Jouni very much for sharing his highly interesting research, and—despite all his duties as director of the Finnish Institute for Educational Research and national project manager of PISA—finding the time for several meetings and fruitful discussions, his great support, and his insights about the PISA assessments.

My sincere thanks to the reviewers Prof. Dr. Agathe Merceron and Prof. Dr. Petri Nokelainen for careful reading of my work and for providing such positive and constructive feedback. It is an honor to have Prof. Dr. Rolf Olsen as opponent. His own dissertation (published more than a decade ago) served as an inspiration and motivation for me.

I want to thank the dean Prof. Dr. Pekka Neittaanmäki very much for his great encouragement and for giving me my first job at this university. Thanks to Dr. Kari Nissinen and Prof. Dr. Antero Malin from FIER for answering all my questions concerning PISA data. I thank Prof. Dr. Mykola Pechenizkiy from TU/e for many interesting discussion about educational data mining.

Prof. Dr. Bülent Yener and Prof. Dr. Mohammed Zaki have helped me very much during my research visit at the data science center at RPI (USA). Thank you both so much for interesting discussions, collaboration, and data mining support! Bülent has also been an incredible host and friend. Besides him I want to thank Michele, Tim, Tony, and Martin for introducing to me to the crazy American life.

I am grateful to my other collaborators, Prof. Dr. Tuomo Rossi, Joonas Hämäläinen, Tommi Lahtonen, Mariia Gavriushenko, Susanne Jauhiainen, and all other colleagues who have supported my research. Thanks to my girls, Elena, Eloho, Yue, and Di for sharing the ups and downs of combining academic with private life.

I acknowledge the PhD program of the Faculty of Information Technology, COMAS, and the Department of Mathematical Informational Technology for generously funding my PhD research. Moreover, I thank these funding parties for financial support to allow me to participate in several conferences and the Science Council of the University for funding my research visit at RPI. Additional support was provided by the Foundation of Nokia Corporation.

Finally, my biggest thanks go to my friends scattered around the world and to my two big families—the one in Germany and the one in Finland—for all their support, love, and encouragement throughout my life. This would never have been possible without you!

GLOSSARY

ACER	Australian Council for Educational Research
CS	Computer Science
DM	Data Mining
DMIT	Department of Mathematical Information Technology
DT	Decision Tree
EAP	Expected A Posteriori
EDM	Educational Data Mining
EDS	Educational Data Science
ESCS	PISA Index of Educational, Social, and Cultural Status
ICT	Information and Communication Technology
IPP	Impact Per Paper
IRT	Item Response Theory
IT	Information Technology
JuFo	Julkaisusfoorumi (Publication Forum in Finnish)
KD	Knowledge Discovery
KDD	Knowledge Discovery from Databases
LA	Learning Analytics
LDA	Linear Discriminant Analysis
LSEA	Large Scale Educational Assessment
MAR	Missing At Random
MCAR	Missing Completely At Random
ML	Machine Learning
MLP	Multilayered Perceptron
MNAR	Missing Not At Random
MOEC	Ministry of Education and Culture
MOOC	Massive Open Online Courses
NAEP	(United States') National Assessment of Educational Progress
NB	Naïve Bayes
OECD	Organisation for Economic Co-operation and Development
PCA	Principal Component Analysis
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PV	Plausible Value
SJR	Scimago Journal Rank
SNIP	Source Normalized Impact per Paper
SoLAR	Society for Learning Analytics Research
SOR	Sequential Overrelaxation
SVM	Support Vector Machines
TIMSS	Trends in Mathematics and Science Study
WLE	Weighted Likelihood Estimates

LIST OF FIGURES

LIST OF TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

PI Mirka Saarela, Tommi Kärkkäinen. Analysing Student Performance Using Sparse Data of Core Bachelor Courses. *Journal of Educational Data Mining, 7(1):3–32*, 2015.

PII Mirka Saarela, Tommi Kärkkäinen. Knowledge Discovery from the Programme for International Student Assessment. *Book chapter in Learning Analytics: Fundaments, Applications, and Trends. Springer, pp. 229–267*, 2017.

PIII Mirka Saarela, Tommi Kärkkäinen. Discovering Gender-Specific Knowledge from Finnish Basic Education Using PISA Scale Indices. *Proc. of the 7th International Conference on Educational Data Mining, pp. 60–67*, 2014.

PIV Mirka Saarela, Tommi Kärkkäinen. Weighted Clustering of Sparse Educational Data. *Proc. of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 337–342*, 2015.

PV Mirka Saarela, Tommi Kärkkäinen. Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. *Proc. of the 8th International Conference on Educational Data Mining, pp. 156–163*, 2015.

PVI Mirka Saarela, Joonas Hämäläinen, Tommi Kärkkäinen. Feature Ranking of Large, Robust, and Weighted Clustering Result. *Proc. of the 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 96–109*, 2017.

PVII Tommi Kärkkäinen, Mirka Saarela. Robust Principal Component Analysis of Data with Missing Values. *Proc. of the 11th International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 140–154*, 2015.

PVIII Mirka Saarela, Bülent Yener, Mohammed Zaki, Tommi Kärkkäinen. Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach. *Machine Learning for Digital Education and Assessment Systems Workshop of the 33rd International Conference on Machine Learning, pp. 1–8*, 2016.

PIX Mariia Gavriushenko, Mirka Saarela, Tommi Kärkkäinen. Supporting Institutional Awareness and Academic Advising Using Clustered Study Profiles. *Proc. of the 9th International Conference on Computer Supported Education, Volume 1, pp. 35–46*, 2017.

PX Mirka Saarela, Tommi Kärkkäinen, Tommi Lahtonen, Tuomo Rossi. Expert-Based versus Citation-Based Ranking of Scholarly and Scientific Publication Channels. *Journal of Informetrics, 10(3):693–718*, 2016.

1 INTRODUCTION

This thesis discusses, in general terms, knowledge discovery from large and sparse educational data through the application and further development of data mining and machine learning techniques and, more specifically, the operation and employment of these methods in the context of the Finnish educational system, which has received a lot of attention during the 21st century. Section 1.1 provides the background and motives for this research, Section 1.2 gives a brief literature review on the Finnish educational system, Section 1.3 poses the research questions, and Section 1.4 explains the structure of this thesis.

1.1 Background and research motivations

Education seems to be the key for a richer and more satisfying life. According to statistics by the Organisation for Economic Co-operation and Development (OECD), higher-educated adults are not only better paid and less likely to become unemployed, they are also more likely to show greater social engagement[1] and to live a longer and happier life (OECD, 2012a). Thus, the average education level of its citizens has a major impact on the general well-being of a country.

Because of the known impacts of education on the general well-being of a population, the interest in international large-scale educational assessment (LSEA) studies and the performance of national education systems are active areas of research. One of the largest and most well-known international LSEAs is the Programme for International Student Assessment (PISA). PISA assesses students' learning outcomes in reading, mathematics, and scientific literacy triennially and is referred to as the "world's premier yardstick for evaluating the quality, equity and efficiency of school systems" (OECD, 2013b).

Finland, a Nordic welfare state with a rather scarce population, has historically performed very well in the international comparison of the PISA assess-

[1] Similarly, research has shown a positive relationship between general intelligence and ethical thinking (e.g., Tirri and Nokelainen, 2012a,b, and references therein).

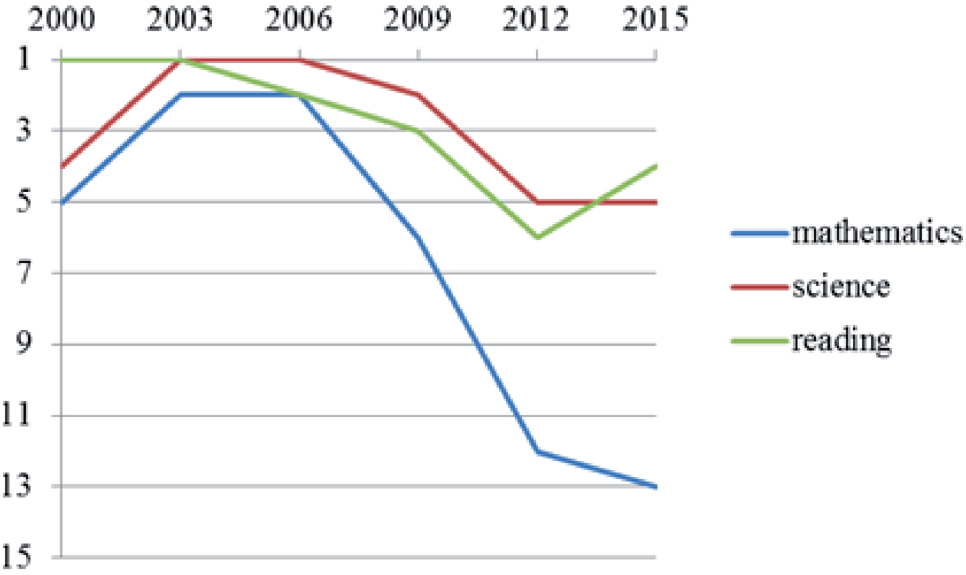


FIGURE 1 Finland’s ranking in the different PISA cycles.

ments and received a lot of praise for its educational system. Especially, the outstanding results in the first three assessments—PISA 2000, 2003, and 2006 (see Figure 1)—made Finland worldwide famous for its high-quality basic education system (Orlowski, 2017; Sahlberg, 2015; Simola, 2014). But also the secondary and higher education have received international attention. According to Schwab et al. (2013), Finland has the world’s best tertiary education, and in a recent BBC press release, Coughlan (2016) argued that a high school degree in Finland is worth more than a tertiary degree in Italy, Spain, or Greece. Especially motivated by the very high results in the first PISA assessments, the unique characteristics of the Finnish educational system have been under active research (see Section 1.2). However, this research has been conducted to date mostly using classical quantitative and qualitative research methods by educational scholars.

The principal aim of this dissertation is to augment the methodological landscape of the LSEA studies, most prominently PISA, to the direction of educational data mining and learning analytics. This is accomplished by developing and applying large scale computational methods arising from pattern recognition, machine learning, data mining, and neural computation to educational data sets with special characteristics. LSEA data are of high quality and publicly available.[2] However, as pointed out by Rutkowski and Rutkowski (2010), scholars often hesitate in using these data because of the many technical complexities within them (this will be further elaborated in Section 3.3 of this thesis). Thus, the contributions of this thesis are twofold and interrelated: first, to advance existing educational data mining and learning analytics methods and algorithms to make them feasible for the big and complicated LSEA data sets with their specific

[2] For example, the PISA data can be downloaded from http://www.oecd.org/pisa/pisaproducts/.

characteristics and, second, to use these methods and algorithms to discover new educational domain knowledge.

According to the PISA 2012 and PISA 2015 results, the performance of Finnish students has strongly declined, most notably in mathematics (see Figure 1), and Finland no longer seems to be the wonderland of educational results. Therefore, the Finnish subset of PISA and the 2012 assessment—where, as illustrated in Figure 1, Finland's large decline in the international ranking was recorded and where the main assessment domain was mathematics—are of main interest in this thesis. The other main data sources utilized in this dissertation are the data from the Finnish publication channel quality ranking system and the study record data from the students of the Department of Mathematical Information Technology at the University of Jyväskylä in Finland. All of these analyzed data sets are extremely sparse.

1.2 The Finnish educational system

When the first PISA results were published, Finland's high ranking came as a surprise to the whole world—especially to the Finns themselves (Sahlberg, 2011). Asian countries also performed very well in the PISA tests and have, particularly in the latest assessments, occupied the highest positions in the country rankings. However, while the Asian education systems are known for extremely long study hours and systematic testing from an early age (Tan, 2017; Liu and Xie, 2014; Waldow et al., 2014; Chua, 2011), the Finnish students have one of the lowest instruction times within the OECD (Reinikainen, 2012), an extremely late start of official school training (i.e., systematic teaching of reading, writing, and mathematics; see **PII**), almost no standardized testing (Simola, 2014; Sahlberg, 2011; Linnakylä et al., 2011; OECD, 2011), and the lowest number of after-school study hours worldwide (see Figure 2). Moreover, the monetary investment in education per student is only average in the international comparison (OECD, 2017b), which seems to make the Finnish system even more phenomenal. Thus, the Finnish system has been praised as a "miracle" (Niemi et al., 2016, 2012; Sahlberg, 2015) and aroused international interest. According to Heller Sahlgren (2015), it even served as a "poster child for many education experts and policymakers throughout the world."

Generally, the Finnish education system partitions into early childhood education, preschool education, basic education, upper secondary education, and higher education. The preschool (age 6), and basic education (age 7–16) are mandatory for all children, with no division into different tracks. The first division takes place in upper secondary education, which students can obtain either at a general upper secondary school or at a vocational institution (MOEC, 2012). Education at a general upper secondary school ends with a matriculation examination that enables enrollment into higher education. Higher education can be obtained either at universities, where the emphasis is on research, or at the more

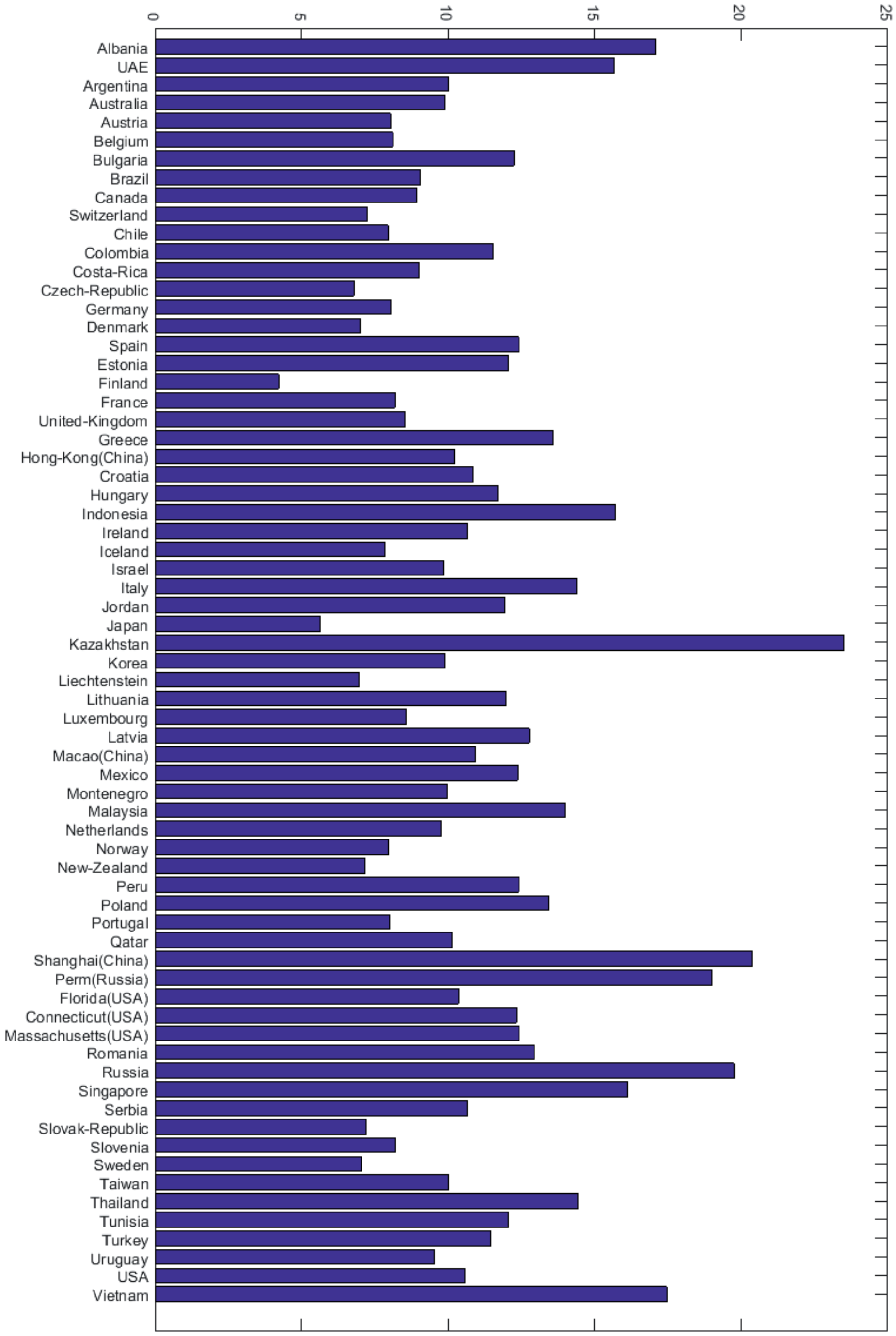


FIGURE 2 Out-of-school study hours for all in PISA 2012 participating countries. In comparison to all the other countries, Finnish students study the least after school. This figure was originally published by Saarela and Kärkkäinen (2017).

vocational oriented polytechnics (MOEC, 2012). Both institutions offer bachelor and master's degrees. However, licentiate or doctorate degrees can be obtained at universities only.

According to a recent review by Schatz et al. (2016), three main reasons attempting to explain Finland's miracle education system can be found throughout the educational research literature. The first reason is the equality and equity in the Finnish school system. As explained above, Finnish pupils are visiting common comprehensive schools from grade 1–9. That means that they are not divided already at an early age into different tracks based on their performance. This inclusion applies not only to students with different performance levels but also to students with special needs (Kivirauma and Ruoho, 2007). Moreover, schools in Finland are publicly funded and offer free learning material, school meals, health care, and school transport for all students (OECD, 2011; Linnakylä et al., 2011). The equality and equity of the comprehensive schools are also shown throughout all PISA cycles. For example, according to the 2012 assessment, the between-school variation in Finland was only 6% of the overall mathematics performance, which is the second lowest figure in comparison with all PISA countries (OECD, 2013a, p.47).[3]

The second reason highlighted in the literature is the autonomy and freedom of educational decision makers. It has been reasoned that instead of market-oriented schooling, standardization of schools and tests (therefore focusing on measurable performance), and competition between students and schools, the focus in Finnish schools is more on cooperation and collaboration (Simola, 2005; Sahlberg, 2011). National curricula, as well as explicit learning objectives and standards exist, but schools and teachers in Finland can decide how to implement learning strategies and pedagogical methods to reach the joint educational goals (Linnakylä et al., 2011; OECD, 2011). Thus, the autonomy and freedom of educational decision makers are higher in Finland than in other countries with more market-oriented education systems.

The third reason usually given for the remarkable Finnish education system relates to the highly qualified teachers, who are appreciated and trusted by the community. Becoming a teacher is a career wish for many young Finns, and only one out of ten applicants gets a place in the primary teacher training programs (Sahlberg, 2011; Linnakylä et al., 2011; OECD, 2011). That means that only the best and most motivated students can become teachers which is one of the reasons Finnish teachers enjoy a very high status in the society (Morgan, 2014; Sahlberg, 2011; Linnakylä et al., 2011; OECD, 2011; Reinikainen, 2012). Traditionally, it has also been reported that parents, and this could be a consequence of and a reason for the highly qualified teachers, are very satisfied with their schools, teachers, and teaching assessment methods (Räty et al., 1995; Simola, 2005).

[3] However, Gaber et al. (2012) argued that precisely the fact that all 15-year-olds are still visiting the common comprehensive school that accommodate students of all levels is the reason for the small between-school difference in Finland, in comparison to, for example, Slovenia, where most of the 15-year-old students were just recently divided based on their performance levels.

As emphasized by Välijärvi et al. (2002) and Välijärvi et al. (2007), these reasons and explanations are mutually interdependent and interrelated—not only with each other but also with the Finnish culture in general. Culture affects peoples' behaviors and attitudes (DiMaggio, 1997; Hitlin and Piliavin, 2004, see also **PIV**). The Finnish culture is often described as a highly collaborative one. For example, the *Hofstede Model* (Hofstede, 2011) characterizes Finland's society as highly "feminine," meaning that the most important driving factors in life are more to care for others and to live a good life instead of focusing on success and wanting to be the best. Simola (2005) traces this less market-oriented culture to some extent back to Finland's geographical location as a borderland to the East that was for some time even was obliged to be part of the Russian Empire. According to him, "Eastern elements are evident in Finland everywhere and in every way, from its administrative traditions to its genetic heredity." Thus, the Finnish society seems to be coined more by appreciation of collaboration, equality, and equity than competitiveness in general, that is, not only from the educational perspective.

1.3 Research questions

In accordance with the goals of this thesis, that is, to advance automatic knowledge discovery methods for sparse educational data and to utilize these methods to learn from the data and to discover new educational domain knowledge, it seeks to answer the following research questions:

RQ1: To what extent can the educational knowledge discovery process be automated?

RQ2: What are the forms and characteristics of the data mining methods and algorithms needed for knowledge discovery from LSEAs, such as PISA?

RQ3: What novel and useful knowledge can be discovered from existing Finnish educational data?

1.4 Structure of the thesis

Table 1 shows how each paper contributes to the research questions and what is the generalizability/impact of the results based on the data. The remaining part of this collection-of-publications-based thesis is organized as follows. In Chapter 2, the overall knowledge discovery process and the emerging research disciplines concerned with knowledge discovery from educational data are described. Chapter 3 depicts the data of the included studies with an emphasis on LSEA data, which are the major data source of the included publications (see Table 1). Chapter 4 provides an overview of the concepts and methods needed for understanding the techniques of the included studies. Then, the included publications and main results are summarized in Chapter 5. Finally, Chapter 6 answers and discusses the research questions posed in Section 1.3, together with their overall implications and conclusions of this thesis, and presents directions for future work. Throughout this dissertation, the included publications are cited in bold using the letter P followed by the Roman numeral of the publication. Reprints of all original publications are attached at the end of this thesis.

TABLE 1 Contribution of original research articles to the research questions.

Article	**Research Question(s)**	**Data (Observations and Variables)**	**Data Source Form/ Collection Method**	**Design**
PI	**RQ 1–3**	13,640 study records with 21 attributes from 1,040 university students	Historical log file data from the university study records	Longitudinal (8/2009-7/2013)
PII	**RQ 1–3**	Global 15-year-old population of 24,720,720 students (weighted from 485,490 student records) with 27 attributes from 68 countries obtained from the PISA 2012 contextual student database	Two-stage stratified sampling (first stage: schools in which 15-year-old students are enrolled, second stage: students in the sampled schools), data stored in big public PISA databases	Cross-sectional and observational
PIII	**RQ 1–3**	Finnish subset (8,829 records) of PISA 2012 contextual student database with 15 attributes	Two-stage stratified sampling	Cross-sectional and observational
PIV	**RQ 1,2**	Population of 60,047 15-year-old Finnish students (weighted from 8,829 records) of PISA 2012 contextual student data with 15 attributes	Two-stage stratified sampling	Cross-sectional and observational

Continued on next page

Table 1 – continued from previous page

Article	Research Question(s)	Data (Observations and Variables)	Data Source Form/ Collection Method	Design
PV	**RQ 1,2**	Same observations with weights as in **PII** with 15 attributes from 68 countries	Two-stage stratified sampling	Cross-sectional and observational
PVI	**RQ 1,2**	Same observations with weights as in **PII** with 38 (15 main and 23 meta) attributes from 68 countries	Two-stage stratified sampling	Cross-sectional and observational
PVII	**RQ 1,2**	485,490 global student records of PISA 2012 contextual student data with 15 attributes	Two-stage stratified sampling	Cross-sectional and observational
PVIII	**RQ 1,2**	Same observations as in **PVII** with 161 attributes (53 from the contextual and 108 from the cognitive PISA 2012 database)	Two-stage stratified sampling	Cross-sectional and observational
PIX	**RQ 1,2**	15,370 study records with 21 attributes from 1,163 university students	Historical log file data from the university study records	Longitudinal (2012-2015)
PX	**RQ 1–3**	29,443 publication channel records with 33 attributes with link to 331,553 records of Finnish publication activities in these channels	Data scraping to extract data of the Finnish publication channels, Finnish publication activity, and Thomson Reuters' Journal Citation Reports database	Longitudinal (2010-2015)

2 TOWARD EDUCATIONAL DATA SCIENCE

The publications constituting this thesis follow the educational knowledge discovery process. This chapter describes the general knowledge discovery process (Section 2.1) and the emerging research disciplines of educational data mining and learning analytics that employ this process for the educational domain (Section 2.2). Following Piety et al. (2014), it is argued that these disciplines may be summarized by the term *educational data science*.

2.1 The knowledge discovery process

Big data, data mining, and data science have become buzzwords in recent years. Data are collected from all kinds of devices and applications, such as web services, satellites, health applications, cars, social media sites, cameras, microphones, smartphones and smartwatches, home appliances, and search engines. New devices are connected to the Internet constantly (see, e.g., Byrne et al., 2017; Chin and Callaghan, 2013; Atzori et al., 2010) and, therefore, produce data collected by service providers and stored in huge databases. These data that originate from such heterogeneous sources are often referred to as *big data*. They are very large and complex and are difficult to analyze with traditional techniques. Data mining algorithms and big data analytics have been developed to understand the collected data, detect unsuspected behavior or events, discover new knowledge from the application domain, and generally learn from the data. These algorithms and analytics have been successfully utilized in various fields, such as business, engineering, social media, and biological science (see, e.g., Chu, 2014). Thus, as Siemens (2014) summarizes these observations, the "message is clear: we live in a world of data and our future promises even greater emphasis on analytics to understand data."

While the big data analytics of the above-mentioned domains steadily grow in importance, the domain of learning and education has historically been rather poor in terms of analytics in comparison (Dawson et al., 2014). Nevertheless,

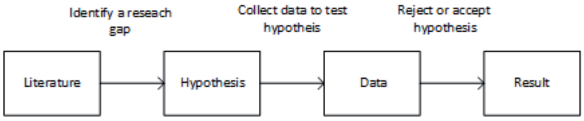


FIGURE 3 Usual steps in the traditional (quantitative) research.

nowadays learning occurs increasingly online, and new virtual learning environments with novel cognitive and collaboration tools have continually emerged (see, e.g. van Leeuwen et al., 2015; Looi et al., 2009, and references therein). Thus, also education has also become an important and growing deliverer of big data. Simultaneously with the increasing masses of available educational data, the research methods have also widened, and more analysis opportunities have emerged (Hershkovitz et al., 2016; Joksimović et al., 2016; Nokelainen and Silander, 2014).

The traditional quantitative research process (see, e.g. Singleton et al., 1993; Creswell, 1994) is roughly illustrated in Figure 3. The researcher usually starts with a hypothesis as part of a research gap from the literature. Then, the researcher designs and develops data collection instruments and collects data that enable the test of the formed hypothesis. Finally, the researcher tests the collected data against the hypothesis so that the hypothesis can be either confirmed or rejected. Thus, a certain set of data is usually manually collected, driven by the existing body of knowledge and developed hypotheses to be assessed and confirmed in the research process. However, as pointed out by Fayyad et al. (1996a), this kind of research process is "slow, expensive, and highly subjective" and became "completely impractical" with the increasing masses of available data from various domains.

In contrast to the traditional hypothesis-driven manual quantitative data analysis approach stands the so-called *data mining* or *secondary data analysis*. In these kind of analyses, the researcher will probably not have been involved in the collection, but usually works with a (large) amount of data that has been collected already for some (other) purpose (Hand et al., 2001; Bryman, 2004a; Breiman, 2003). Data are the center of data mining research. According to the definition by Hand et al. (2001), one of the key features of data mining is that the interesting and useful patterns in the data should emerge automatically without the need to form a strong hypothesis first. This is probably the most distinguishing feature from data mining to the traditional data analysis. Thus, this kind of research has an exploratory nature (Tukey, 1977), but it does not rule out hypothesis testing and confirmatory research to properly assess novel findings.

A huge bulk of information is gathered and available in data, and powerful algorithms exist that could provide insights into the data domain as well as uncover its hidden patterns. However, the availability of big data and big data analytics does not eliminate or replace the researcher. On the contrary, studying the target domain and theories is still one of the most important and substantial parts of research (Merceron et al., 2016). Moreover, real-world data sets seldom come in a format that immediately allow the data mining procedure. They might

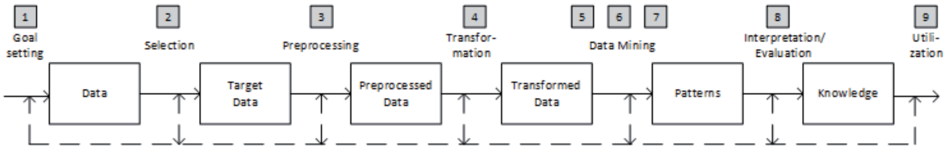


FIGURE 4 The knowledge discovery process according to Fayyad et al. (1996a). Generally, this process involves the nine steps marked by the small gray quads.

be noisy and very sparse. Furthermore, existing algorithms may need adaption to customize to the domain and the problem at hand. Another challenge is that no definition exists that specifies what are interesting and useful patterns. An analogy for this is like finding a needle in a haystack—without even knowing the appearance of the needle, or if a needle is hidden in the haystack at all. Hence, a knowledge discovery process includes many interactions from the human side, which to a great extent involve and require a profound understanding and knowledge of the studied domain.

The overall *knowledge discovery (from databases)*[1] process was introduced by Fayyad et al. (1996a,c,b) in several articles and is illustrated in Figure 4. As shown in the figure, the actual data mining is just a part of the broader knowledge discovery process (Zaki and Meira, 2014). This process is not specific to any particular application domain. It starts with a thorough study of the target domain and its masses of available data and is intended to deliver useful knowledge to be utilized in the respective target domain as the end product. The general knowledge discovery process mainly consists of the nine steps (the gray quads in Figure 4) that are briefly illuminated below. Moreover, as illustrated by the dashed lines in the figure, if the results of a step are not satisfactory, it is possible to repeat single steps or to go back to a preceding step.

1. *Learning the target domain and setting a goal:* As emphasized above, the first step of the knowledge discovery process requires a profound study of the particular application domain and its theories involved, and the available data. Based on this target domain study, a reasonable knowledge discovery goal is set.
2. *Selection:* A suitable subset of data is selected, and a target dataset is created. This is an important step because the different ways data are collected and the selection of reduced parts of the original data can affect the data analysis results considerably. Naturally, analysis results or discovered knowledge are not generalizable if the data are not representative of the studied phenomenon (Pelánek et al., 2016).
3. *Preprocessing:* In the third step of the knowledge discovery process, the selected data is cleaned, verified, and preprocessed. This involves operations such as removing noise or outliers, deciding how to deal with missing data, and removing of duplicate or unreliable data. Moreover, additional vari-

[1] In the data mining literature, the terms *knowledge discovery from databases* (KDD) and *knowledge discovery* (KD) are used interchangeably. In this thesis, only the latter term will be utilized.

ables might be defined, constructed, or derived from the existing ones. This step is often referred to as the most laborious one that consumes most of time and the bulk of effort invested in the data analysis (e.g., Vesanto et al., 1999; Gaber et al., 2005; Tan et al., 2007; Witten et al., 2011). Many data mining algorithms are readily implemented and accessible (for example, in Phyton, Matlab, or Weka) but no implementation can perform the preprocessing automatically for any possible data set. The many different ways data can be collected and stored is only one reason for this.

4. *Data transformation:* The preprocessed data are transformed to make them feasible for the data mining procedure. This step may involve conversions of variable types (for example, categorization or binarization) but also dimension reduction methods, such as feature selection and extraction. Feature selection is the process of selecting those features that contain the most important information and neglecting the remaining features, while feature extracting methods extract the most important information of all the original features and transform them into a more compact set of new features (Alpaydin, 2010).
5. *Data mining method selection:* The data mining method that will be applied to the transformed data is decided. Different taxonomies of data mining methods exist in the data mining literature, but the most common division is into predictive or supervised (for example, classification, and regression), descriptive or unsupervised (for example, clustering), and patterns and rules discovery methods (Hand et al., 2001; Tan et al., 2007; Han et al., 2011; Bramer, 2007; Zaki and Meira, 2014). The method of the data mining step depends on the knowledge discovery goal and on the availability and form of the data. For example, supervised methods can be used only if the target output (for example, class labels in classification) is available.
6. *Data mining algorithm selection:* Several different algorithms exist for each of the data mining categories listed above. In the sixth step of knowledge discovery, the particular data mining algorithm is decided together with its parameter settings.
7. *Data mining:* The selected algorithm is applied to the data. Depending on the data mining method, the end product of this step can consist of numerical results, such as clusters, association rules, or classifiers.
8. *Interpretation and evaluation:* The numerical results obtained from the mining step are converted into a useful and understandable form. Visualization techniques are used to illustrate the obtained structures and models. Patterns and/or visual models extracted from the original data are presented in such a way that domain experts or other involved parties can easily understand and use the discovered knowledge.
9. *Utilization:* The final step of the process involves usage activities of the discovered knowledge from the previous step. This knowledge might, for example, be incorporated in systems of the application domain, used for quantified decision making, and/or documented and made (openly) available for interested parties.

2.2 Educational data mining and learning analytics

Educational data mining (EDM) and learning analytics (LA) are emerging (Gray et al., 2014) research fields at the interface of educational data sets and computational data analysis methods. Application of these methods (or data analysis in an educational context, generally) typically realizes an *educational knowledge discovery process* (**PII**), following the common knowledge discovery process illuminated in Section 2.1 for the specific domain of education. EDM emerged roughly in 2004 and became a fast growing research line, with its first annual conference held in 2008 and the advent of the Journal of Educational Data Mining in 2009. LA emerged around the same time as EDM (Piety et al., 2014), but its own publishing and presenting forums were established a little bit later. The first International Learning Analytics & Knowledge Conference was held in 2010, and the Journal of Learning Analytics was established in 2014. The Society for Learning Analytics Research was founded in 2011 (Ferguson et al., 2015).

EDM is concerned with the development of methods for exploring, understanding, and benefiting from data that come from educational settings (Romero and Ventura, 2010). It is defined as “an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (International Educational Data Mining Society, 2016). Thus, it consists of developing or utilizing data mining methods that are especially feasible for discovering novel knowledge originating in educational settings (Baker and Yacef, 2009) and supporting decision making in educational institutions (Calders and Pechenizkiy, 2012). Most of EDM case studies analyze the steadily growing amount of log data from different computer-based learning environments, such as learning management systems (for example, Valsamidis et al., 2012), intelligent tutoring systems (for example, Hawkins et al., 2013; Bouchet et al., 2012; Carlson et al., 2013; Springer et al., 2013), or educational games (for example, Kerr and Chung, 2012; Harpstead et al., 2013).

In comparison, LA is defined as a discipline to “measure, collect, analyze, and report data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens, 2013; Ferguson, 2012). It aims for discovery and communication of meaningful and actionable patterns in educational data (Pardo and Teasley, 2014; Gray et al., 2014; Siemens and Baker, 2012) and the whole range of factors that affect learning, including the learners’ inner and outer actions and their learning environment (Peña-Ayala, 2017). This is often accomplished by visualizing these factors in distinct dashboards for interested parties, such as the learner, teachers, headmasters, lectures, leaders of educational institutions, and so on (Ferguson and Shum, 2012; Verbert et al., 2013). Thus, LA primarily attempts to improve learning and the educational environment by raising awareness, while EDM has a slightly more technical focus.

Although EDM and LA may hold their own specific scopes, as discussed above, they are interdisciplinary by nature and there has been some confusion about strictly categorizing studies in these fields. The principal aim in both fields is to discover novel or unsuspected and useful information from educational data. To discover such information from educational data, Baker (2010) classified EDM methods into five categories:

1. prediction,
2. clustering,
3. relationship mining,
4. discovery with models, and
5. distillation of data for human judgment.

Chatti et al. (2012) stated that LA techniques for detecting interesting educational patterns originate from

1. statistics,
2. information visualization,
3. data mining (equalizing this with the knowledge discovery process illuminated in Section 2.1 and further subcategorizing this category into prediction, clustering, and association rule mining), and
4. social network analysis.

These classifications indicate that the methods in LA and EDM are also very cognate and interrelated. This applies in particular to the different data mining techniques listed in both taxonomies, as also illustrated in Figure 5. Although these techniques constitute only one category in Chatti et al.'s taxonomy, they are generally acknowledged as the fastest growing (e.g., Siemens, 2013) and the most sophisticated (e.g., Rogers, 2015) LA methods. In fact, with the increasing sizes of preexisting educational data, LA scholars tend to shift from utilizing more traditional data analysis techniques, such as statistics, to the more scalable data mining methods (Ferguson, 2012; Hershkovitz et al., 2016; Joksimović et al., 2016, see also the discussion in **PII**).

Siemens (2013) had already anticipated that the LA and EDM disciplines would overlap soon, and only a year later, Piety et al. (2014) stated that the distinctions between EDM and LA "were blurry from the start and in recent years have converged." In fact, Dawson et al. (2014) found that the most influential studies (determining those through citation analysis, see **PX**) in the EDM and LA field are those that try to define these specific research areas. As these disciplines have become increasingly mature, empirical EDM and LA studies have grown in their importance (Dawson et al., 2014). Nevertheless, in summary, these observations underline the difficulty in defining clear boundaries between EDM and LA.

Many scholars describe EDM and LA as "complementary" (e.g., Gray et al., 2014; Siemens and Baker, 2012) or "sister" (e.g., Siemens, 2014; Baker and Inventado, 2014) research fields. The notion that EDM and LA are related complementary disciplines is also supported by the fact that most members of the Society for

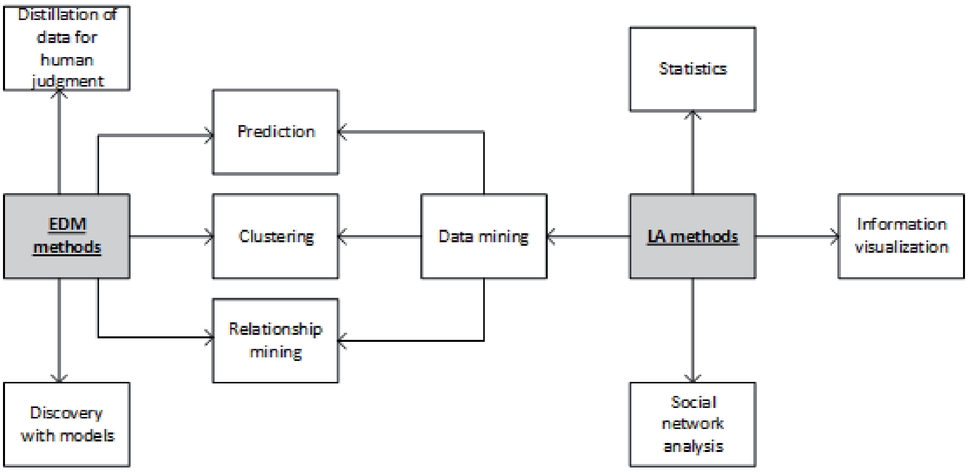


FIGURE 5 Relation between the taxonomies of the EDM and LA methods. The emphasis of the methodology in this thesis lies in the intersection of these two taxonomies. Educational data science methods cover all the categories contained in the figure.

Learning Analytics Research, that is, LA scholars, belong to the EDM society too and vice versa and that they typically visit the same conferences (Siemens and Baker, 2012; Baker and Inventado, 2014). Therefore, instead of categorizing the EDM and LA research fields into mutually exclusive groups, Piety et al. (2014) proposed using the broader term *educational data science*. This terminology will also be used in this thesis to generally refer to the analysis of large existing data sets originating from educational settings through the application and advancement of big educational data analytics and data mining algorithms. In fact, the methodology underlying this thesis focuses exactly on the intersection of EDM and LA methods, as illustrated in Figure 5, that is, prediction, clustering, and relationship mining in the educational domain. Educational data science methods are thus meant to cover all the categories contained in Figure 5, with the focal point on this intersection.

3 ON SPARSE LARGE-SCALE EDUCATIONAL DATA AND THEIR DOMAIN

In this thesis, data mining techniques are utilized and developed to discover knowledge from big and sparse data originating from educational settings (see Figure 5). This chapter defines sparsity (Section 3.1) and big data in the educational domain (Section 3.2). By following the first part of the educational knowledge discovery process (as described in Section 2.1), it then describes the educational target domains and selected data that were analyzed in the included publications. The main data source is from the 2012 PISA cycle (Section 3.3). The other data sources utilized in this dissertation are the data related to the management of a national educational system, that is, the Finnish publication channel quality ranking system (Section 3.4) and the study record data from the students of the Department of Mathematical Information Technology at the University of Jyväskylä (Section 3.5).

3.1 Sparse data and types of missing values

Assume that a set of observations $\{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbf{R}^n$, is given so that N denotes the number of observations and n the number of variables, respectively. If the data matrix $\mathbf{X} \in \mathbf{R}^{N\times n}$ is defined as $\mathbf{X} = \left(\mathbf{x}_i^T\right), i = 1, \dots, N$, a matrix $\boldsymbol{P} \in \mathbf{R}^{N\times n}$ can be defined to indicate the availability of the values in $\boldsymbol{X}$ with

$$(\boldsymbol{p}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases} \tag{1}$$

Throughout this thesis, $\boldsymbol{X}$, that is, the matrix with missing and non-missing/observed values, will be referred to as the *whole* or *full* data. As in Allison (2002), the process of finding values for missing data in $\boldsymbol{X}$ will also be called *completing* the data. If $\boldsymbol{X}$ does not have any missing values, that is, all entries in $\boldsymbol{P}$ are one, $\boldsymbol{X}$ will be referred to as *complete* data. Moreover, if there are missing data in $\boldsymbol{X}$, that subset of observations that has non-missing values for all variables will be referred to as the *complete* data (subset). Note that the whole/full and complete data are the same when all observations have observed values for all variables.

Sparse data are data with many missing values. Missing values can occur for a number of reasons. They occur especially in the face of high-dimensional data because data points are located in a larger space when the number of dimensions increases (Chen et al., 2009; Verleysen and Francois, 2005; Dash and Liu, 2000). This is also called the *curse of dimensionality*. The analysis of data with missing values requires special methods. These methods typically depend on the type of missing data. Rubin (1976) and Little and Rubin (2002) have set the guideline for classifying missing data in the statistical data analysis. They distinguish three types of missing data:

1. Missing completely at random (MCAR), if the probability that the jth component of a vector $\boldsymbol{x}_i$ is missing is independent of any other known or missing values of $\boldsymbol{x}_i$, that is, the missingness occurs completely at random and is independent both of observable variables and of unobservable parameters of interest. In this case, any missing data treatment can be used without introducing bias to the data (Batista and Monard, 2003). However, in practice, the MCAR is seldom present. One example where MCAR might occur in a real data set is when the data are *missing by design* (Allison, 2002), that is, for example, when part of a questionnaire data is missing because part of the questionnaire was not administered to some participants.
2. Missing at random (MAR), if the probability that the jth component of a vector $\boldsymbol{x}_i$ is missing does not depend on the values of missing components, but may depend on the values of observed components, that is, the missingness occurs not completely at random, but can be fully accounted for by variables where there is complete information.
3. Missing not at random (MNAR), if the probability that the jth component of a vector $\boldsymbol{x}_i$ is missing depends on its value, that is, the value of the variable which is missing is related to the reason it is missing. In the statistical analysis, this is "the least desirable situation" (von Davier, 2014, p. 178). However, as pointed out by Pyle (1999), and as can be seen in **PX**, this situation can lead to important information in the knowledge discovery process.

These three notations are also used in this thesis to refer to the type of missing data.

3.2 Big data in education and large-scale educational databases

Generally, *big data* is defined by four Vs, where the first three go back to Laney (2001), and the last one has been added among others, for example, by Gupta et al. (2014):

- *Volume* refers to the size of data sets caused by the number of data points, their dimensionality, or both;
- *Velocity* is linked to the speed of data accumulation;
- *Variety* stands for heterogeneous data formats, which are caused by distributed data sources, highly varying data gathering, and so on; and
- *Veracity* refers to the fact that (secondary) data quality can vary significantly, and manual curation is typically impossible.

Baker (2015) restricts this definition somewhat for the area of education. He characterizes "big data in education" as big by comparison to most classical education research, but "not human genome project or Google big." As pointed out in **PII**, big data in education became a research focus most notably with the advent of massive open online courses (MOOCs). MOOCs are distance learning courses that are made available through the Internet and, often free of charge, accessible by anyone with computer and Internet access. Because of the large and heterogeneous population of MOOC participants, the analysis of MOOC data provides many opportunities (see, e.g., Wang et al., 2014; Ye and Biswas, 2014; Reich et al., 2014) but also a lot challenges, especially because of the large amount of missing values in them (Bergner et al., 2015b,a).

Other examples of big educational data are data from LSEAs (see, e.g., **PII**, **PVI** and **PVIII**). The first LSEAs had emerged already in 1960 (Waldow et al., 2014). They include, for example, the United States' National Assessment of Educational Progress (NAEP),[1] the European Survey on Language Competences (ESLC),[2] the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS).[3] All of these LSEAs assess educational achievements on a large scale, and except the NAEP, even beyond national boundaries. Although all LSEAs share a common focus on measuring educational achievements, they differ in their objectives (Wagemaker, 2014), scope, and the assessed student population. For example, while PISA assesses educational achievements of 15-year-old students, TIMSS is conducted for fourth and eighth graders. An overview of some international LSEAs, their scope, number of participating countries, and timing can be found in Table 3.1 of the work by Heyneman and Lee (2013). Similarly, Appendix 1.A in the very recent book by Lietz et al. (2017) provides a comparison of which countries have participated in which LSEAs. This comparison shows that PISA covers the most countries of all analyzed LSEAs.

[1] nces.ed.gov/nationsreportcard/

[2] www.surveylang.org/

[3] See both http://timssandpirls.bc.edu/

3.3 The Programme for International Student Assessment

Domain

As emphasized in the introduction (Section 1), Finland's educational system became famous because of the high results in the first PISA assessments. PISA is a worldwide study by the OECD and one of the largest (see Section 3.2 above) and most politically acknowledged LSEAs (Knodel et al., 2014). It not only assesses the reading, mathematics, and science proficiencies of students in different countries but also provides data about "learners and their contexts" (see **PII**), such as the students' demographic data and their attitudes and behaviors toward various aspects of education. Thus, Schleicher (2007) asserts that PISA provides "one of the most powerful predictors for the success of an education system."

The target population of the PISA assessments are all 15-year-old students enrolled in a school within the participating countries. In Finland, virtually all students of this population attend the common comprehensive school, whose main features were highlighted and described in Section 1.2. As also pointed out in Section 1.1, the utmost educational knowledge discovery interest here is the success and the recent rapid decline of the Finnish 15-year-old student population in mathematics in PISA. Hence, the focus is on the 2012 assessment, where the rapid decline in mathematics was recorded the first time and where mathematics was the main assessment domain.

Data

According to the OECD, PISA results have a high degree of validity and reliability (see, for example, OECD, 2009, 2012b) so that they can be used to assess and compare the educational systems of the participating countries. Therefore, PISA data should not suffer from the selection bias described in Section 2.1, and analysis results should be representative for the whole assessed student population, that is, all 15-year-old students enrolled in a school within the participating countries. Although PISA data are of high quality and publicly available, little research has been conducted on the secondary analysis of PISA data. According to Olsen (2005a), large amounts of money[4] are spent on ensuring the quality related to the development of the PISA data collection instruments, procedures of PISA data collection, and storage of PISA data in public databases. However, less money is invested into the analysis of these data, and a large set of information in the PISA data is typically not analyzed at all as part of the primary agenda of these assessments (Olsen, 2005a). This is interesting given that almost all data are easily accessible for researchers in public databases.

Rutkowski et al. (2010) argued that the sizes of PISA data sets as well as the technical complexities within them—that is, the sparsity, the weights, and

[4] For example, only in Germany, the costs of the PISA assessment so far aggregate to 21.5 million euro (Musik, 2016).

the many variables derived through complex models—may be the reason only a few researchers work with this freely available and high-quality data (see also the discussion in **PII**). Commonly, those scholars interested in the PISA domain are educational researchers who may be familiar with the traditional research approaches (see Section 2.1), but not necessarily with the intricacy of the big and complicated PISA data sets. They may in fact be completely unable to read or write using programming languages. Hopmann et al. (2007), for example, stated that the PISA technical reports are incomprehensible for anyone who is not personally involved in the implementation of the PISA studies him- or herself, and Spiegelhalter (2013) referred the PISA methods simply as "opaque."

That the PISA data are not trivial can also be concluded from the time that is needed for the analysis: PISA results are published usually approximately 1.5 years after the data collection (e.g., the PISA 2012 data collection took place in spring 2012, and the results were published at the third of December 2013). Moreover, PISA data are big data in education (see Section 3.2). For example, the student questionnaire data set for 2012 alone already consisted of 485,490 observations and 634 variables.[5] A further example is the documentation of the technical steps behind PISA: The 390-pages-long *PISA 2009 Technical Report* (OECD, 2012b) was published in 2012, and the 472-pages-long technical report for the assessment in 2012 was published in the end of 2014 (OECD, 2014b).

To enable the knowledge discovery from these sparse educational data sets, special methods are needed. As pointed out above, PISA data have certain characteristics that have to be taken into account when working with them. First, all available PISA data sets are very sparse. Most of the missing values in PISA are missing by design (see Section 3.1). For each student who participates in PISA, the time is limited to two hours for the cognitive test and to half an hour for a background questionnaire. However, the total cognitive assessment material developed for PISA exceeds these 120 minutes of testing time by far so that each student is administered only a fraction of the entire item battery. For example, the 2012 cognitive test consisted of an item battery that was 450 minutes long. This arrangement of the test is called *rotated design* (OECD, 2014b) or *multiple-matrix sampling* (Rutkowski et al., 2016; Rutkowski, 2014, 2011).

Because of this rotated design, 74% of the data are missing from the data set, which includes the scores of the cognitive test for the single test items. Moreover, since the 2012 assessment, this rotated design has also been applied to the background questionnaire to increase the total quantity of contextual information that can be assessed (Adams et al., 2013). More precisely, three different background questionnaires were included in the 2012 assessment, and only one was administered to each student. All three versions included a common part about the student and his or her family and home, but the remaining parts were varied.[6] Because of this arrangement, many variables in the contextual data set have about 30% missing data (see, e.g., **PII**, **PIII**, **PIV**, and **PV**).

[5] See http://pisa2012.acer.edu.au/downloads/M_stu_codebook.pdf

[6] The different questionnaires of the 2012 assessment are available at https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm.

Second, PISA—and LSEAs in general—include sampling weights (see, e.g., Meinck, 2015; Rutkowski et al., 2010). Only a fraction of 15-year-old students from each country take part in the assessment. However, when multiplied with their respective weights, they should represent the whole student population. For example, the sample data of the 2012 PISA study consisted of $485,490$ students that, taking the weights into account, represented more than 24 million 15-year-old students in the 68 different countries and territories that participated in PISA in 2012. To select a reliable sample of the 15-year-old student population, the OECD applies a two-stage sampling design in each country: First, schools attended by 15-year-old students are assigned to mutually exclusive groups based on explicit strata, and schools from these groups are selected with probabilities proportional to their size. Then, students within those school are selected randomly with equal probability.

The real-valued weight w_i assigned to each participating student i consists of the school base weight, the within-school base weight, and five adjustment factors, especially the one that compensates the non-participation of a sampled student (OECD, 2014b, see also **PVI**). Hereby, both over- and under-sampling has taken place in PISA for various student groups, such as the deliberate over-sampling of immigrants as well as students from Swedish-speaking schools in the Finnish subset of the PISA 2012 assessment. Thus, it is important to utilize the weights at each stage of the analysis (see in particular **PIV**, **PV** and **PVI**) to achieve unbiased population estimates and to report findings that are valid for the whole population. This also means that the weighted averages of the national samples are used when the populations of the participating countries are compared, such as in Figure 2.

Third, many variables in PISA data sets are derived variables, that is, not the raw assessment data. Most of these variables have been constructed using the item response theory, one of the most fundamental paradigms in psychometrics. Therefore, to be able to work with PISA data, one should also have an understanding of how the many derived variables have been created and how they can be used for further analysis. In fact, integrating methods from the psychometrics literature with methods from the machine learning and data mining literature is a common characteristic of many knowledge discovery studies in the educational domain (see the discussion by Baker, 2010).

As part of the standard PISA preprocessing phase, certain scale indices are constructed based on information gathered from the background questionnaire for each participating student (OECD, 2013b). These indices describe, for example, students' engagement, drive, and self-beliefs, and are constructed using item response theory. Furthermore, as described above, 74% of the scored cognitive item data are missing. That means that proficiencies in PISA are not observed directly, but must be inferred from the available sparse scored item response data and the students' background obtained from the contextual questionnaire. For this, the so-called plausible value technology is utilized, which will be further elaborated below (Section 3.3.1).

To address the above-mentioned challenges, the publications included in this thesis discuss and advance techniques that enable the knowledge discovery process from these distinguished LSEA data by taking their specific characteristics, that is, the sparsity, the weights, and the many derived variables, into account.

3.3.1 Plausible values and measuring performance in PISA

As described above, 74% of the PISA 2012 scored cognitive test data are missing. That means that national group and subgroup proficiencies in PISA are not directly observed, but must be inferred from the available sparse scored item response data and the students' background obtained from the contextual questionnaire. For this, a posterior distribution of the range of abilities that a student might reasonably have is estimated (OECD, 2014b), given his or her answers in the background questionnaire and the sparse observed item responses of the cognitive test. The plausible values that are reported in the PISA data (similarly as in all LSEAs listed in Section 3.2) are simply random draws from the estimated posterior distribution. This technique was originally developed by Mislevy (1991) for the NAEP LSEA (see Section 3.2) and is based on Rubin's (1987) work on multiple imputations. Thus, student performances in PISA are not observed but completely imputed (von Davier, 2014, p.184).

The posterior distribution of student abilities, which the plausible values are drawn from, is estimated with Bayesian statistics. According to Bayes' theorem (see, for example, 1.44 in Bishop, 2006), the *posterior distribution* is proportional to the product of the *likelihood* and the *prior distribution*. Hence, the posterior distribution of a student's ability can be modeled as follows:

$$f(\beta \mid \boldsymbol{x_i}, \boldsymbol{y_i}) \propto P(\boldsymbol{x_i} \mid \beta, \delta) f(\beta \mid \boldsymbol{\lambda}, \boldsymbol{y_i}), \tag{2}$$

where $P(\boldsymbol{x_i} \mid \beta, \delta)$ denotes a Rasch model (explained in **PII**) given the student's ability β and the difficulties δ of the items in the test administered to the student, and $f(\beta \mid \boldsymbol{\lambda}, \boldsymbol{y_i})$ denotes a population model. The *prior distribution* is a population model that is estimated with a latent regression model, with $\boldsymbol{\lambda}$ denoting the regression coefficients and $\boldsymbol{y_i}$ denoting collateral variables in the data (for example, variables modeling background information of the student i). This latent regression model estimates the average proficiencies of examinee subgroups, given evidence about the distribution and associations of the collateral variables (Marsman, 2014; von Davier and Sinharay, 2013; Wu, 2005).

In PISA 2012, the collateral variables included in the latent regression model were “all available student-level information, other than their responses to the items in the booklets" (OECD, 2014b, page 157). The *likelihood* of success in the cognitive test is a Rasch model, where the probability of success is a kind of logistic function of the latent ability (unknown, will be estimated) and some parameters (e.g. difficulties) of the test items. The Rasch model assumes that a student with high ability should give correct answers to items with higher probability than a student with low ability, and that a student should give a correct answer to an easy item with higher probability than to a difficult item. The obtained *posterior distribution* of a student's ability is specific for each student, since each student has different values of background variables and test results. In the PISA data, five plausible values (for each domain) drawn from this posterior distribution are reported for each student.

To sum up, student proficiencies in PISA are not directly observed, and it is important note that the plausible values are estimates for group performance and should never be used as test scores of individual examinees (OECD, 2014b; Von Davier et al., 2009). They are a selection of likely proficiencies that could have been observed if the student had taken the whole test of all cognitive items (Adams et al., 2013; Wu, 2005).

3.3.2 Related work on PISA cluster analysis

To the knowledge of the author of this thesis, clustering of PISA data has been performed only in four other studies in addition to the studies included in this dissertation. Both Kjærnsli and Lie (2004) and Olsen (2005b) used hierarchical clustering to cluster the residual matrix of science items. These residual matrices were created in the following way: For each country and each science item, the percentage of correct items was computed, and then the average over countries for a particular item and the average over items for a particular country was calculated. The residual matrix shows how much better or worse a certain country scores on a certain item. Kjærnsli and Lie (2004) performed the clustering using the residual matrix from PISA 2000 data and Olsen (2005b) using the residual matrix from PISA 2003 data. Clustering of contextual PISA data was performed only in the publications included in this thesis, that is, **PII**, **PIII**, **PIV**, and **PV**, and two recent master's theses[7] (Koskela, 2016; Wallden, 2016).

3.4 The Finnish publication channel quality ranking system

Domain

Teaching and research are the two fundamental activities in higher education institutes, such as universities. Research is generally evaluated by codifying and disseminating newly produced knowledge in the form of publications in scientific publication channels (Abramo et al., 2016). To translate these publications into an evaluation system, one has witnessed a transition from the raw numbers of different kinds of publications (e.g., books, articles, reports) toward their aggregated quality indicators (Haustein and Larivière, 2015), which has become an important constituent in national resource allocation models of higher education institutes in many countries (Auranen and Nieminen, 2010; Fairclough and Thelwall, 2015). Various performance-based funding systems are currently in action

[7] Both theses were supervised by the author of this thesis and her first supervisor.

and under continuous evaluation and development in different countries (e.g., Hicks, 2012; Wilsdon et al., 2015; Kulczycki and Rozkosz, 2017).

In Finland, the publication activity of an individual university has been part of the national funding instrument since 2007. Initially it was based on a rough categorization of publications together with direct aggregation. However, together with the fundamental renewal of the university legislation in 2010, Finland has renewed its university resource allocation systems to introduce a component that aggregates the quality and quantity of publications by following the Norwegian model (see **PX**) implemented in Norway and Denmark. This means that since then, the quality of an individual publication has been taken into account through quality ranks (0–3) and through the corresponding weighting factors for the overall publication productivity.

Currently, 13% of public funding for a national university is based on the aggregated ranks of all the publications that were produced over three years. The main driver for creating a unified national ranking system for all relevant publication channels were, especially, the difficulties in using the available quality measures over all the disciplines (research and publication culture, e.g., in humanities and social sciences as compared to that of technology and natural sciences). The purpose of the National Publication Forum, *JuFo*,[8] is to recognize all relevant publication sources, that is, series and publishers, in order to pinpoint to the national scientific community the characteristics of various places to publish. The national aim, naturally, is to target research activity to prestigious international forums and also to enable national evaluation and management of research activities and its quality over the years. Hence, *JuFo* serves in Finland as both an indicator of the quality of publication channels and as a guideline to allocate funding to the national HEIs.

Generally, the quality of a publication channel can be evaluated based on either (i) the judgment of an expert in the area (expert-based) or (ii) citation-based indicators of scientific impact (Ahlgren and Waltman, 2014; Ahlgren et al., 2012). Citation-based indicators judge the quality of a publication channel according to measures of citations. Although highly cited publications do not always indicate impactful research, this premise tends to be true on average (Waltman et al., 2013). Thus, publication channels of articles with a high number of citations can be considered to be of higher quality than publication channels with low citation rates (White, 1990). When the quality of a publication channel is determined by a group of specialists, the evaluation is called expert-based.

Currently, the classifications in *JuFo*, that is, the Finnish ranks, are expert-based. Each publication channel is assigned to exactly one of altogether 24 different expert panels. The first 23 expert panels represent different scientific areas (all areas can be found in Table A.12 in **PIX**) and are composed of experienced and respected Finnish researchers in these areas. The one remaining panel is responsible for evaluating interdisciplinary sources. Moreover, a steering group provides the common rules for ranks and rankings, most importantly the portion of the highest ranks at the levels 2–3.

[8] JuFo is the abbreviation of 'Julkaisusfoorumi,' which means Publication Forum in Finnish.

The expert panels must classify all the publication channels assigned to them into one of the four quality categories (0-3), in such a way that level 3 should represent the top, level 2 the leading, and level 1 the basic scientific publication channels in the respective panel area. Level 0 is for those channels that either do not meet the basic requirements, such as those being fully peer reviewed and having an editorial board constituted by experts, or that have not yet been evaluated because the channel has been admitted just recently to the *JuFo* list. To contribute to a university's funding, the channel of a publication must be on the *JuFo* list. Moreover, the percentage of publication channels that a panel is allowed to classify as "leading" or "top" are restricted to 20% (and 5% respectively). Thus, only a very small percentage of the publication channels of all disciplines can receive the highest, and in terms of funding, the most valuable level rating.

The rank levels of all publication channels in the *JuFo* list are reevaluated every fourth year. In addition, new publication channels are admitted to the list four times a year. During these intermediate 'complementary evaluations' level 0 publication channels can be upgraded to level 1, and level 1 publication channels can be downgraded to level 0. As these decisions are made by well-respected (and therefore generally highly paid) researchers, they are expensive in terms of labor, administration, and salary.

Data

In 2016, *JuFo* incorporated almost 30,000 different publication channels with 33 attributes. Besides the Finnish expert-based rank, the Norwegian and Danish expert-based rankings are also incorporated into *JuFo*. Moreover, the three main citation-based indicators from the bibliographic database Scopus are featured, and by using the ISSN linkage to Thomson Reuters' Journal Citation Reports, the eight citation-based indicators stored there can be accessed for the common publication channels. Furthermore, other variables that might affect the ranking of publication channel (such as the age and the discipline) are provided in the *JuFo* database, and through a link, one can directly access the information of all researchers in Finland who have published in the particular channel. Since all publication channels are missing some of the featured attributes in *JuFo*, and not all publication channels are listed in all bibliometric databases, one again faces a significant sparsity problem. All of this is explained in more detail in **PX**.

3.5 Student records from the Department of Mathematical Information Technology

Domain

Besides research and teaching (see Section 3.4), the staff in Finnish universities is also obliged to perform certain administrative tasks. Such tasks can include

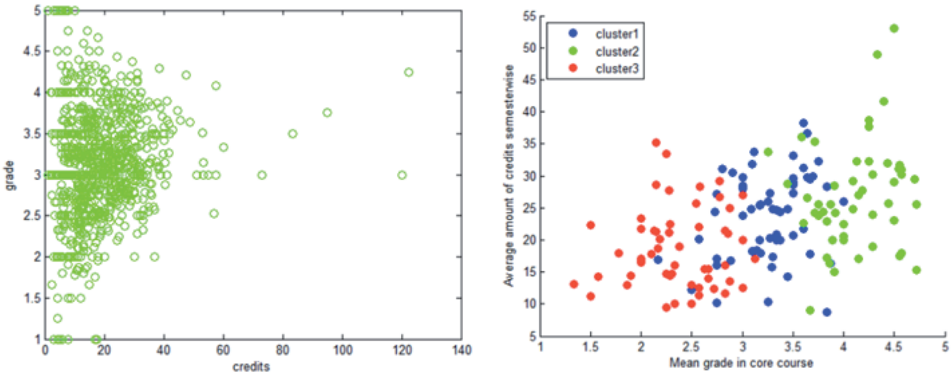


FIGURE 6 Mean credits compared to the mean grade of all studies of the DMIT students (left) and mean credits and the mean grade of the DMIT mandatory bachelor courses only (right). The figures show that while the general mean credits versus the mean grade assemble the classical bell curve and are not correlated ($r = 0.0848$), the mean credits versus the mean grade for the mandatory courses are positively correlated ($r = 0.4415$) These figures were originally published by Saarela and Kärkkäinen (2015).

curricula and personal study plan creations or recommendations, which can be challenging and time and labor consuming, particularly when performed for all students individually. This applies especially to the computer science field because of the universally known high dropout rates in this discipline (e.g., Kinnunen et al., 2013).[9] Also, the Department of Mathematical Information Technology (DMIT) at the University of Jyväskylä, which is comparable to a computer science program at other universities, records high dropout and low study progression rates of its students, particularly when compared to the students of the other department of this university (this is explained in more detail in **PI**). Hence, there is an intensive need of academic advising at DMIT, and the question is which skills a student should possess or which skills should be promoted and supported to ensure that the students progress and succeed in their studies.

Data

A data warehouse of passed courses by all the students of the university exists and is available to authorized parties. That means that a real study path can be studied to enhance part of the manual administrative work. Each study record is complete, featuring attributes of the student (such as his or her name, gender, and birth date), the passed course (such as the title, unique identifier, and the number of credits), and the grade the student obtained in this course. However, when these records are transformed into a matrix where the students are the observations and the courses the variables, this matrix becomes extremely sparse, with not a single complete observation (**PI**).

[9] See also the ranking by The Daily Telegraph (2017) for or a more recent documentation of this observation.

If all courses were included in the matrix, the sparsity pattern would clearly not be treatable anymore, as students have the opportunity to choose from a large pool of courses and therefore accumulate study records of very different courses (**PI**). When the focus is only on that set of courses that is mandatory for all DMIT students (and thus, should be the most complete), one still faces a sparsity problem, but this one is significantly less severe. Moreover, as can be seen from the left side of Figure 6, the general mean credits completed by DMIT students versus their mean grade assemble the classical bell curve and are not correlated at all. That means that students who complete many courses are not necessarily also more successful in terms of grades. However, the mean credits versus the mean grade for the mandatory courses are positively correlated (right-side of Figure 6). Thus, the subset of all mandatory bachelor courses provide an informative, and in terms of sparsity, still treatable part of the data.

4 FOUNDATIONS OF CONCEPTS AND METHODS

This chapter introduces the basics concepts and methods that are needed for understanding the utilized and further developed analysis techniques in the included publications. Since all publications deal with sparse data first, a short overview of standard procedures to deal with missing data is provided (Section 4.1). Then, the spatial median as a robust location estimate is discussed (Section 4.2), since the unsupervised methods of many of the included publications are based on or utilize this estimate to deal with the sparsity in the analyzed data. Finally, those unsupervised (Section 4.3), supervised (Section 4.4), and frequent pattern mining (Section 4.5) data mining algorithms that are of importance for a comprehension of the articles are briefly addressed.

4.1 Standard procedures to deal with missing data

As pointed out above, LSEA and real-world educational data sets are often very sparse. This sparsity might be conditioned by design (such as in PISA, see Section 3.3), by nature (such as in JuFo, see Section 3.4) or through preprocessing and transformation (such as in the DMIT records, see Section 3.5). Concerning big data (see Section 3.2), sparse data can cause the last V, that is, low veracity. Little and Rubin (2002), who, as discussed in Section 3.1, set the guideline for classifying missing data, divide the strategies to analyze such data into four, mutually not exclusive, categories: (i) strategies that remove observations that have missing values, (ii) weighting strategies, (iii) strategies that impute missing values, and (iv) model-based strategies. This classification scheme is often considered the golden standard for the treatment of data with missing values in the traditional data analysis (von Davier and Sinharay, 2013).

Strategies from the first category, which Little and Rubin (2002) call *procedures based on completely recorded units*, delete all observations that have at least one missing value. The advantage of this strategy is its simplicity and straightforwardness. It is also often the default solution in software implementations of data mining algorithms, such as in the Matlab's `k-means` implementation. The disadvantage of this strategy is that part of the data will be lost. This strategy is thus especially unsuitable for data with a high percentage of missing data.

Strategies from the second category, *weighting procedures* (Little and Rubin, 2002), modify the sampling weights as if the missing data were part of the sample design. This is accomplished by assigning each complete observation a weight so that the weighted complete data sample distribution approximately fits the real population. The advantage of weighting procedures is that they are nonparametric in the sense that they only require a model for the available data probabilities (which have to be estimated from the data, for example, through logistic regression), but not for the data values in the population. Moreover, they are easy to apply for univariate data with monotone missing data patterns. However, for multivariate data with an arbitrary missing data pattern, weighting procedures are not recommended, as a different set of weights may have to be computed for each feature (Schafer and Graham, 2002).

Strategies from the third category, *imputation-based procedures* (Little and Rubin, 2002), use the data values that are present in the data set to estimate the value(s) of the missing entry/entries. Various imputation-based procedures exist. The *mean-imputation* substitutes all missing values with the mean value of the not missing data. While this procedure is relatively simple, it underestimates the variance of the data, as the mean by definition does not contribute to the variance. The *hot deck imputation* compares each observation with missing data to all observations with complete data and chooses the most similar complete observation as a "donor" for the observation with missing data. One way of doing this was presented in **PI**: Cluster all data and use the centroid of the cluster to which the observation with missing values was assigned as its donor. The *cold deck imputation* works similar to the hot deck imputation. However, while in the hot deck imputation the donor is taken from the same data set for which the missing value is imputed, in the cold deck imputation, the donor is taken from another data source (Batista and Monard, 2003). The *prediction-based imputation* estimates missing values by predicting them with the help of regression or classification models based on available data of the other attributes.

Multiple imputation imputes missing values M times (usually three or five times) using an appropriate model (Rubin, 1987; Allison, 2002). Then, the desired analysis (e.g., linear regression) is performed on each of the M complete data sets. Finally, the parameter estimates (for example, the coefficients and standard errors) obtained from each analyzed data set are combined for interference.

Strategies from the last category, which Little and Rubin (2002) denominate as *model-based procedures*, estimate the parameters of a model defined for the complete data. One popular example for a model-based procedure is the *maximum likelihood* estimation. However, such estimations are based on the assumption that the data of interest come from a certain family of distributions.

In a more recent article, Cheema (2014) reviewed missing data treatment methods specifically for the educational domain. He distinguished only *deletion* and *imputation* methods. As discussed above, deletion methods are seldom recommended as missing data treatment and are especially unsuitable for data with many observations of missing data. For example, in the case of PISA 2012, where because of the rotated design less than 30% of the questionnaire scale indices data are complete,[1] only a very tiny subset of the data would be kept. While in the PISA cognitive data, where because of the design of the test none of the observations are complete (**PVIII**), there would not be any data left at all.

The main disadvantage of the simpler imputation-based procedures is that they make assumptions on data density and/or reduce variance. However, despite the known problems that arise from these missing data treatments, Cheema (2014) found that many educational researchers prefer to continue using these simple methods to handle sparse data.

To sum up, traditional techniques for handling sparse (educational) data seem to focus on either using only that subset of the data that is completely available, and therefore, reducing the size of the original data set or making assumption about the missing data and concerning the unknown density distributions. In comparison, the idea of robust clustering (Äyrämö, 2006), that is described below, and the principal component analysis technique developed in **PVII** is to take *all available data*, as defined in (1), into account without making any assumptions concerning the unknown density distributions of the data. This means that nothing known is discarded and nothing unknown is included.

4.2 The spatial median as location estimate

The most established statistical family of probability distributions is the *normal* or *Gaussian* family (Sprent and Smeeton, 2007). A normal distribution $\mathcal{N}$ is defined by the mean μ and the variance σ^2. If the mean and the variance of a normally distributed data set are known, the probability of the values can be computed easily for any new observation from this data set. This classical statistics based on μ, σ^2, and least-squares-error are also referred to as *second-order statistics*. Most of the traditional data analysis methods are based on second-order statistics.

Moreover, when a statistical analysis is carried out, it is often assumed that the data is a random sample from the normal or another parametric distribution, such as the binomial, multinomial, Poisson, or exponential distribution. However, in many real-world applications, it is unreasonable to assume that a data sample comes from a certain parametric distribution. In such cases, it is preferable to use so-called nonparametric statistical methods that are not based on such strong assumptions on the underlying distribution (Huber, 2009; Sprent and Smeeton, 2007).

The sample mean is the most efficient estimator[2] for the samples that are

[1] In the Finnish subset, 2,520 of the 8,829 observation (see **PIII**) and 142,394 of the 485,490 observations in the whole PISA 2012 (see **PV**) have complete data for the 15 scale indices.

[2] In statistics, the most efficient estimate of a comparison is considered the one with the lowest variance (Everitt and Skrondal, 2002).

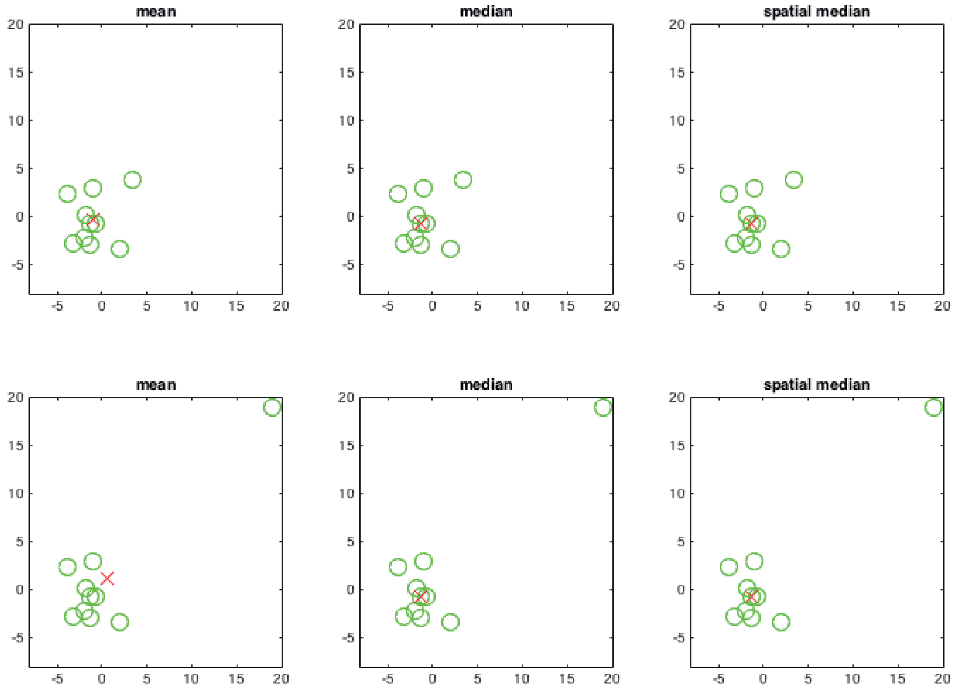


FIGURE 7 Example of location estimates of a data set without outliers (first row) and their change when one observation of this data set is transformed to an outlier (second row). Data observations are illustrated as green o's and the location estimates as red x's. The example illustrates that the sample mean is extremely sensitive toward the one outlier, while the median and spatial median remain almost the same.

drawn from the normal distribution. But since real-world data rarely satisfy this normal assumption (Kontkanen et al., 2000) and for a nonsymmetric or skewed distribution, estimators other than the sample mean may be preferable. In particular, the sample mean is highly sensitive to all kinds of outliers and for that reason very non-robust (Huber, 2009). This is shown in Figure 7. As argued, for example, in **PI**, **PIII**, and **PVII**, a missing value can, in principle, represent any value from the possible range of an individual variable so that it becomes difficult to justify assumptions on data or error normality. Since all data sets analyzed in this thesis are very sparse, instead of focusing only on the second-order parametric and non-robust statistics, the *nonparametric first-order robust statistics* are also considered. First-order robust statistics allow deviations from normality assumptions while still producing reliable and well-defined estimators (Rousseeuw and Leroy, 1987; Huber, 2009, **PVII**).

The two simplest robust estimates of location are median and spatial median. The median, a middle value of the ordered univariate sample, is inherently one-dimensional, and with missing data uses similarly to the mean, only the available values of an individual variable from the marginal distribution. The spatial median, however, is truly a multidimensional location estimate and utilizes the available data pattern as a whole. Mathematically, the spatial median is the point that minimizes the sum of the Euclidean distances to a group of points $\{\mathbf{x}_i\}_{i=1}^{M}$ (for example, the points assigned to the same cluster). It can be formulated as

$$\underset{\boldsymbol{c}\in\mathbf{R}^n}{\arg\min}\,\mathcal{J}(c),\text{ for }\mathcal{J}(c)=\sum_{i=1}^{M}\|\boldsymbol{c}-\boldsymbol{x}_i\|_2. \tag{3}$$

Although the basic concept of this point is easily understood and has been extensively discussed in the literature, albeit under various names[3] (Drezner and Hamacher, 2001), its computation is known to be difficult.

In the univariate case, (3) is equivalent to the coordinatewise sample median, that is, when the values $\boldsymbol{x}_i$ with $i = 1 \dots M$ are sorted in increasing order, the value at position $\frac{M+1}{2}$ if M is odd, and the whole interval of middle values if M is even (see, e.g., Kärkkäinen and Heikkola, 2004). Moreover, for $\boldsymbol{c} \neq \boldsymbol{x}_i$, the gradient is unique and straightforward to compute. However, for $\boldsymbol{c} = \boldsymbol{x}_i$, the subgradient has to be employed since the absolute value function is non-differentiable at the zero point.

On one hand, the breakdown point of the sample mean is zero (Rousseeuw and Leroy, 1987). The spatial median, on the other hand, is characterized by a breakdown point of 0.5, meaning that it can handle up to 50% of the contaminated data. Moreover, as shown by Kärkkäinen and Heikkola (2004), the cost function of the spatial median depends only on the directions and not on the magnitudes of $\boldsymbol{c} - \boldsymbol{x}_i$, which considerably decreases the sensitivity toward outliers, especially compared to the sample mean (see Figure 8). Thus, the spatial median is an attractive location estimate for high-dimensional data with severe

[3] The spatial median is, for example, also known as the multivariate L1-median or the Fermat-Weber point.

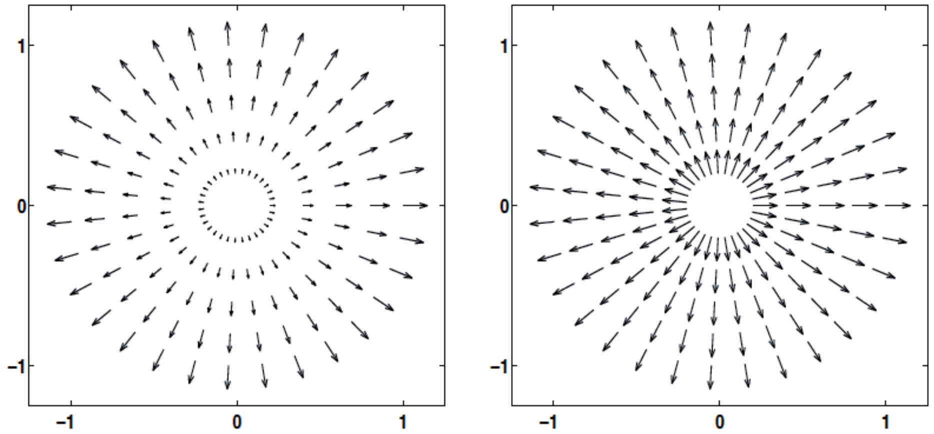


FIGURE 8 Gradient fields of $\|x_2^2\|$ (left) and $\|x_2\|$ (right), where $\|\cdot\|_2$ denotes the l_2-norm of a vector (Kärkkäinen and Heikkola, 2004). The length of the gradient vectors increases for the sample mean that is based on $\|x_2^2\|$. Therefore, the sample mean is very sensitive toward outliers and not a robust location estimate. On the other hand, the spatial median, which is based on $\|x_2\|$, depends only on the direction of the data and gives equal weights for all observations. This shows that the spatial median is a very robust location estimate.

degradations and outliers, possibly in the form of missing values (**PI**, **PIII**, **PIV**, **PV**, **PVI**, **PVII**, and **PIX**).

4.3 Unsupervised methods

Unsupervised data mining and machine learning methods refer to techniques that do not need labels. Clustering is discussed in Section 4.3.1 and principal component analysis is addressed in Section 4.3.2.

4.3.1 Clustering

Clustering as an unsupervised method is the process of dividing points into groups so that the points within one group are similar to each other and the points in different groups are dissimilar to each other (Jain et al., 1999). These different groups of points are called clusters. Jain (2010) defines an ideal cluster as a group of points that is "compact and isolated." Various clustering methods and approaches exist, such as density-based clustering, probabilistic clustering, grid-based clustering, and spectral clustering (Aggarwal and Reddy, 2013), but the classical division of clustering methods is to distinguish *hierarchical* and *partitional* methods (Celebi et al., 2012; Jain, 2010; Merceron and Yacef, 2005; Tan et al., 2007; Celebi and Kingravi, 2012).

Hierarchical clustering methods enable visual summarization of the hierarchies and orders in a given data set through the *dendrogram*. They can be divided into agglomerative and divisive techniques. Agglomerative clustering techniques operate in a bottom-up fashion, that is, they start with each observation as a separate cluster and then repeatedly merge the most similar clusters C_m and C_n so that they form a new bigger cluster. In the case of *Single Link*, the most similar clusters are defined as those that contain the shortest distance,[4] δ, between a point in C_m and a point in C_n,

$$\min\{\delta\left(\boldsymbol{u}, \boldsymbol{v}\right) \mid \boldsymbol{u} \in C_m, \boldsymbol{v} \in C_n\}, \tag{4}$$

while other methods define similarity differently (*Ward's method*, for example, uses the minimum variance). Divisive hierarchical clustering work in the opposite direction of agglomerative techniques, that is, they start with all observations in the same big cluster and then recursively split the most dissimilar clusters until each observation forms its own cluster. However, because of the pairwise distance matrix requirement, hierarchical clustering is not scalable to a large number of observations (Zaki and Meira, 2014, page 372).

Partitional (or representative-based) clustering, on the contrary, is very scalable and efficient even for large data sets (Celebi et al., 2012). Another advantage of these clustering techniques is that they assign each observation to exactly one cluster that is represented by the cluster centroid, that is, the middle point of the cluster. As the middle point, the cluster centroid represents the most common profile of all points within that cluster. This makes partitional clustering very attractive from the knowledge discovery point of view (see Section 2.1), because instead of looking into all the points in a cluster, one can interpret each cluster based on its most representative point (see, for example, **PI**, **PIII**, **PIV**, **PV**, and **PIX**).

From k-means to k-spatial-medians clustering

Generally, partitional-based clustering algorithms consist of an initialization step in which the initial centroids of each cluster are decided and two iterative steps in which (i) each observation is assigned to its closest centroid, and (ii) the centroid of each cluster is recomputed by utilizing all observations assigned to it. The algorithm stops when the centroids remain the same in two successive iterative runs. The most popular and commonly applied partitional-based clustering method is `k-means` (Jain, 2010). The objective of the `k-means` algorithm is to minimize the sum of the squared error over all K clusters,

$$\mathcal{J}(\{\mathbf{c}_k\}_{k=1}^{K}) = \sum_{k=1}^{K}\sum_{i=1}^{M_k}\|\boldsymbol{x}_i - \mathbf{c}_k\|_2^2, \tag{5}$$

[4] Different distance measures might be used; the most common is the Euclidean distance $\delta\left(\boldsymbol{u}, \boldsymbol{v}\right) = \sqrt{\sum_{i=1}^{N}(u_i - v_i)^2}$.

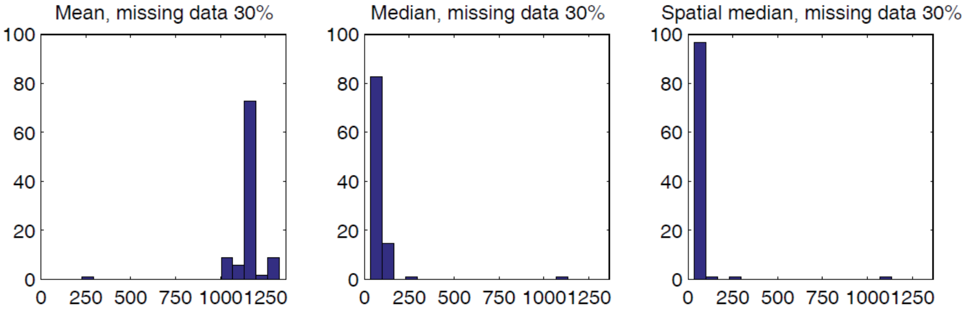


FIGURE 9 Error distributions from 100 test runs on a simulated 2-dimensional data set of 30 observations with 30% of missing data for the three location estimates mean, median, and spatial median (Kärkkäinen and Äyrämö, 2004).

where M_k ($M_k \leq N$ with N being the total number of observations in the data, see Section 3.1) denotes the number of observations attached to a particular cluster k, $k = 1 \ldots K$. The gradient of $\mathcal{J}$ with respect to the kth centroid is given by $\nabla \mathcal{J}(\mathbf{c}_k) = \sum_{i=1}^{M_k} \mathbf{x}_i - \mathbf{c}_k$. Setting this term to zero yields the coordinatewise sample mean of the M_k points in the corresponding cluster k:

$$\sum_{i=1}^{M_k} \boldsymbol{x}_i - \mathbf{c}_k = 0 \Leftrightarrow M_k \mathbf{c}_k = \sum_{i=1}^{M_k} \mathbf{x}_i \Leftrightarrow \mathbf{c}_k = \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbf{x}_i.$$

The `k-means` algorithm works very well for complete and mixed Gaussian data since the sample mean is the most efficient estimator for samples that are drawn from the normal distribution. However, as discussed above in Section 4.2, the sample mean is highly sensitive to all kinds of outliers, as well as missing values, which can be characterized as special types of outliers. In fact, Kärkkäinen and Äyrämö (2004) showed that the `k-means` algorithm produces unreliable results for 10% of missing data and that the quality of the clustering result decreases the more that missing data are introduced. Figure 9 shows that the spatial median still produces very reliable results even when 30% of data are missing, while the sample mean is clearly not feasible anymore for data with such sparsity patterns.

The sensitivity to missing values is a problem and makes the `k-means` unusable for data with a high sparsity pattern. Dash and Liu (2000) point out that also most other clustering algorithms perform poorly for sparse data. The `k-spatial-medians` clustering algorithm, which was introduced by Äyrämö (2006), utilizes the same basic steps as the `k-means`, but the objective function is to minimize the spatial median, that is, the sum of the Euclidean distances to the M_k attached points to the kth cluster. Äyrämö (2006) solved the difficulty of computing the spatial median by using the sequential overrelaxation (SOR) algorithm with the overrelaxation parameter $\omega = 1.5$ (see Äyrämö, 2006, for details). The SOR algorithm introduced by Young (1954) is an iterative method for solving a linear system of equations that accelerates the convergence. In addition, in the implementation of `k-spatial-medians` clustering, only the available data are taken into account when the centroid is recomputed by using the projections as defined in (1). Therefore, the whole objective function reads as follows (note that the Euclidean distances are not squared):

$$\mathcal{J}(\{c_k\}_{k=1}^{K}) = \sum_{k=1}^{K} \sum_{i=1}^{M_k} \|\mathrm{Diag}\{\boldsymbol{p}_i\}(\boldsymbol{x}_i - \boldsymbol{c}_k)\|_2, \tag{6}$$

where Diag transforms a vector into a diagonal matrix and the $\boldsymbol{p}_i$'s indicate the sparsity pattern, that is, the available variables observationwise as defined in (1).

To conclude, the robustness to missing and noisy data and the fact that every cluster is represented by a centroid make `k-spatial-medians` clustering very suitable for knowledge discovery from sparse data. As illustrated in **PI**, **PIII**, **PIV**, **PV**, and **PIX**, each cluster can be interpreted by describing its most representative point, the available data spatial median. Moreover, this clustering algorithm with its available data strategy does not make any assumptions about the underlying distribution of the data or the type of missing data (see Section 3.1), discarding no available information.

Determining the number of clusters

As pointed out above, the goal of clustering is to find groups of points so that the points within one group are similar to each other and dissimilar to the points in the other groups. This similarity between points is usually computed with a distance measure: One wants to obtain a clustering result that has small within-cluster distances but large between-cluster distances. The most common approach to determine the number of cluster K is to look at the clustering error (for example, for `k-means`, the sum of the squared distances of all points to their centroids as defined in (5)), and select that value for K for which the clustering error is reasonably small compared to the number of clusters. In the literature, the plot of the clustering error for different K values is often simply referred to as *knee point* or *elbow curve* (e.g. Thorndike, 1953).

The problem with the elbow curve is that it only looks into the first objective of clustering (minimizing the within-cluster distances), but not into the other objective of maximizing the separation of different clusters. Naturally, the elbow curve decreases for each newly introduced cluster until it becomes zero when each observation is allocated to an individual cluster (see, for example, Figure 3 in **PI**). Different cluster indices have been introduced that make use of both of these cluster objectives. These cluster indices provide thus a more comprehensive way to determine the number of clusters; examples include the Ray-Turi (Ray and Turi, 1999) and the Davies-Bouldin (Davies and Bouldin, 1979), as well as as the Davies-Bouldin* (Kim and Ramakrishna, 2005) indices. Cluster indices have been utilized, compared, and advanced in several papers (see, e.g., **PII**, **PIII**, **PIV**, and **PV**; Jauhiainen and Kärkkäinen, 2017; Arbelaitz et al., 2013; Liu et al., 2010), generally with the conclusion that no index shows advantage over the remaining indices in every context.

Initialization

Partitional iterative clustering algorithms, such as those described above, are very sensitive with regard to their initialization (see also the discussion in **PI**). This means that the clustering result and quality vary depending on the initial choice of the centroids. If the initial centroids are chosen in a less optimal way, the algorithm will probably not converge to the global optimum. Obviously, trying each point separately is not feasible, as this already takes a very long time for a small set of points. In fact, the number of all possible combinations is given by the Stirling number of the second kind,

$$S(N,K) = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\binom{K}{i}i^N. \tag{7}$$

Random initialization is still an often chosen alternative. Celebi and Kingravi (2012) provided an overview of different initialization techniques proposed in the literature. According to them, `k-means++` (Arthur and Vassilvitskii, 2007), where the random initialization is based on a density function favoring distinct centroids, is one of the best initialization methods to date. This is also evidenced by the fact that `k-means++` is currently the initialization method in Matlab's `k-means` implementation. However, when dealing with sparse data there are additional initialization challenges, as the centroids needed for iteration and final interpretation must be complete (**PI**, **PIII**, **PIV**, and **PV**).

Evaluation

Cluster evaluation is difficult and can be very discouraging (Jain and Dubes, 1988). One example for this is that, usually, not just one right solution exists. There are often (many) different solutions that can be equally valid and meaningful (Jain, 2010). Moreover, since establishing the labels is exactly the idea of clustering, there is usually no ground truth or one single right solution given beforehand that can be compared to the clustering result. If such cluster labels were known for all observations, supervised methods could be used.

Jain and Dubes (1988), similarly to Zaki and Meira (2014), distinguished three cluster evaluation approaches: internal, relative, and external.

- *Internal* cluster evaluation refers to measures that can be derived from the data. Examples of such measures include the cluster indices (see above) that compute distances between clusters and distances of observations within clusters to assess the quality of a clustering result (**PIII**, **PIV**, **PV**, and **PIX**). Another example is the Kruskal-Wallis test statistic (Kruskal and Wallis, 1952) that can be used to compare different clusters and assesses whether the observations originate from the same distribution (**PI** and **PVI**).
- *Relative* cluster evaluation refers to directly comparing separate clustering results, often for the same algorithm (Zaki and Meira, 2014), for example, comparing the clustering results when one parameter, for example, the num-

ber of clusters (**PI**, **PIII**, **PIV**, and **PV**) or the initialization (**PI** and **PIV**), is changed.
- *External* cluster evaluation refers to an evaluation where the ground truth is known. This does not necessarily refer to prior given labels, but could also refer to a domain expert who specifies meaningful structures of the data or metadata that can explain the clusters (**PII**, **PIII**, **PIV**, **PV**, and **PVI**).

4.3.2 Principal Component Analysis

Principal component analysis (PCA), also known as the Karhunen-Loéve transformation, is one of the most famous and widely used linear dimension reduction methods (Jolliffe, 2002; Alpaydin, 2010). As a dimension reduction method, PCA is mainly employed in the preprocessing and transformation steps of the knowledge discovery process (Section 2.1). It generates a new set of variables, called principal components in such a way that

- each principal component is a linear combination of the original variables,
- all the principal components are uncorrelated (i.e., orthogonal in dimensional space) to each other, and
- all the principal components are ordered so that the first few retain most of the variance present in all the original variables.

To find the optimal m-dimensional ($m \ll n$) subspace that contains most of the variance of the original n-dimensional data, PCA uses the eigenvectors and corresponding eigenvalues of the covariance matrix Σ,

$$\Sigma = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})\,(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{8}$$

These eigenvectors are ordered by their eigenvalues. The eigenvector with the largest eigenvalue is called the first principal component and indicates the direction of most of the variance in the data.

This can be seen by projecting the data onto a one-dimensional subspace with the direction defined by a vector $\mathbf{u}_1$ (see, for example, Zaki and Meira, 2014; Bishop, 2006). Since only the direction of the maximal variance, and not the magnitude of $\mathbf{u}_1$, is of interest, a constraint can be imposed to the vector $\mathbf{u}_1$ so that $\mathbf{u}_1^T\mathbf{u}_1 = 1$. Each data point $\mathbf{x}_i$ is projected onto $\mathbf{u}_1^T\mathbf{x}_i$ so that the variance of the projected data is

$$\frac{1}{N}\sum_{i=1}^{N} (\mathbf{u}_1^T\mathbf{x}_i - \mathbf{u}_1^T\bar{\mathbf{x}})^2 = \mathbf{u}_1^T\Sigma\mathbf{u}_1. \tag{9}$$

To maximize the projected variance $\mathbf{u}_1^T\Sigma\mathbf{u}_1$ with respect to $\mathbf{u}_1$, a Lagrange multiplier λ is introduced to enforce above constraint $\mathbf{u}_1^T\mathbf{u}_1 = 1$:

$$\mathcal{L}(\mathbf{u}_1, \lambda) = \mathbf{u}_1^T\Sigma\mathbf{u}_1 + \lambda(1 - \mathbf{u}_1^T\mathbf{u}_1). \tag{10}$$

Differentiation of (10) with respect to $\mathbf{u}_1$ yields

$$\frac{d\mathcal{J}}{d\mathbf{u}_1} = 2\Sigma\mathbf{u}_1 - 2\lambda\mathbf{u}_1. \tag{11}$$

Setting (11) to zero yields

$$2\Sigma\mathbf{u}_1 = 2\lambda\mathbf{u}_1 \Leftrightarrow \Sigma\mathbf{u}_1 = \lambda\mathbf{u}_1. \tag{12}$$

This implies that λ is an eigenvalue of the covariance matrix Σ, with the associated eigenvector $\mathbf{u}_1$.

The second principal component indicates the next largest variance and is orthogonal to the first one. This can be repeated until m principal components are selected that explain most of the variance in the data. These first m eigenvectors are used to project the data points to a new coordinate system, where the axis directions are spanned by the eigenvectors and thus contain maximal variance. The remaining principal components can be discarded without losing a lot of information.

This classical PCA is based on the sample covariance matrix and the sample mean (8), which as discussed in Section 4.2 are extremely sensitive toward all kinds of degradations and outliers in the data, including missing data. Thus, the principal components are also very sensitive with regards to sparse data (Äyrämö, 2006). Moreover, the eigenvalues of the standard covariance matrix Σ represent the *variances* along the new coordinate system ($\lambda_k = \sigma_k^2$), which overemphasize the major components. That means that to assess the true *variability* or *spread* of data in one direction (which the geometric interpretation proposes, see **PVIII**), the standard deviation ($\sigma_k = \sqrt{\lambda_k}$) should be used to determine the real importance of a principal component $\mathbf{u}_k$, $k = 1, \ldots, n$ (Bishop, 2006).

4.4 Supervised methods

Supervised methods can be divided into regression methods, where objects are assigned to continuous values, and logistic regression and classification methods, where objects are assigned to one of several predefined categories, the so-called (class) labels or classes. This is accomplished by building a model, that is, the regressor or classifier, that learns for a set of given objects how their input variables relate to their continuous or class label output variables. Once the model has learned the relation, it can be used to automatically predict the output variable from the input variables of any new object.

Because of the model-based dependance of the descriptive variables to the performance variables, that is, the plausible values, in PISA (as described in Section 3.3.1), the focus of the publications included in this thesis is on unsupervised methods, especially clustering. However, in **PVIII**, an approach is presented to derive labels that are based on the actual PISA cognitive test performance only, and these labels are then used to compare different classifiers in their prediction accuracy. Moreover, in the publications concerned with the other two data sets, **PI** and **PX**, supervised prediction techniques are used in the triangulated analyses. Numerous categories of different prediction models exist. Thus, solely the main ideas of the utilized supervised prediction methods are briefly illuminated below.

Probabilistic classifiers use probability theory to find the most likely of the possible classes. *Naïve Bayes* (NB) employs Bayes' theorem to perform the classification. It estimates the joint probability density function for each class (which is modeled as a multivariate normal distribution) and predicts the class that maximizes the posterior probability. To simplify this estimation, it is naïvely assumed that all attributes are independent given the class. In spite of this unrealistic assumption, NB classifiers have shown good results in practice (Domingos and Pazzani, 1997). While NB classifiers assume a multivariate normal distribution, *nearest neighbour* (NN) classifiers are non-parametric, that is, they do not make any assumptions about the underlying joint probability density function. NN classifiers, which were introduced by Fix and Hodges (1951), estimate the class probabilities of a new object using the class(es) of its closest observation(s).

The idea of a *linear discriminant analysis* (LDA) classifier, which was originally introduced by Fisher (1936), is to map the data into a space where the classes are separated the most. It can be compared to PCA (see Section 4.3.2) because both techniques attempt to find linear combinations of attributes that best approximate the data. However, while PCA is an unsupervised technique that finds the axis directions that maximize the variance of the original data, LDA is a supervised method that tries to find that linear combination of the original attributes that maximize the distinction between the class labels. With the help of the so-called kernel trick, LDA classifiers can also be used for nonlinear classification (Mika et al., 1999).

A *support vector machine* (SVM) uses a hyperplane for linear classification. This hyperplane is fitted to the data in such a way that the margin between the classes is maximized. Similarly to LDA classifiers, SVM classifiers can also be used for nonlinear classification when the kernel trick is utilized.

Decision tree (DT) classifiers build a hierarchical tree-like structure where every node represents one attribute test condition and every directed edge represents one decision (Breiman et al., 1984; Quinlan, 2014). At each node, it splits the data so that each partition has a purer distribution of observations from (a) certain classe(es). Thus, a DT model is easy to read and provides understandable rules on the splitting attribute for human interpretation (see Figure 10). A *random forest* first builds several different DTs. Then, it classifies new objects as the mode of the classes of all its individual trees (Breiman, 2001).

Artificial neural networks (ANN) cover a range of different models that correspond to mathematical models inspired by the biological information processing in the brain. A *multilayer perceptron* (MLP) is a particular category of ANNs. Out of all ANN categories, it has shown the best value in practice (Bishop, 2006) and is also the most widely used ANN (Hand et al., 2001). Early ideas for ANNs can

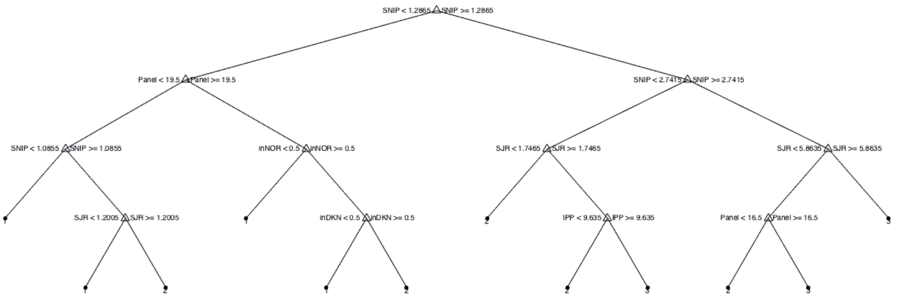


FIGURE 10 Example of a pruned decision tree for the research domain. The *source normalized impact per paper* (SNIP) indicator is the variable with the highest predictive power for the rank in the Finnish funding system. This figure was originally published by Saarela et al. (2016).

be found as early as 1943 in a study by McCulloch and Pitts. Recently, ANNs have become popular again through the concept of *deep learning* (LeCun et al., 2015).

As the desired output is known in supervised methods, it is fairly straightforward to compare different classifiers in their prediction accuracy for a given data set. To assess this accuracy, the data is usually divided into a *training* and a *test set*. Different division strategies exist, but the most established is *cross-validation*. Cross-validation divides the data into N folds and then uses each of the N-folds once for testing and the remaining $N - 1$ folds for training the respective classifier. The overall prediction accuracy of the classifier is then determined as the mean of the N different test accuracies. When the cross-validation is *stratified* (see, for example, **PI**), the folds are created according to some rule. For example, the folds might be created in such a way that in each fold the distribution of observation from the different classes is the same. If there is no rule and the data is, for example, randomly divided into $N = 10$ different folds, the cross-validation is referred as *unstratified*. A *confusion matrix* shows for each class label how many observations were predicted to be from which class label. Thus, the diagonal of a confusion matrix matches the correct predictions, and the remaining part of the matrix matches the false predictions.

4.5 Association rule mining

The goal of frequent pattern mining is to automatically detect interesting and potentially useful patterns in data. In the publications included in this thesis that used frequent pattern mining (**PIII** and **PX**), the goal was to find patterns of strongly associated attribute values, referred as itemsets. However, frequent pattern mining can also be used to detect more complicated patterns, such as sequences or graphs (Zaki and Meira, 2014).

Frequent itemset patterns can be detected with association rule mining. Association rule mining allows researchers to present the discovered patterns as implication rules (Agrawal et al., 1993). If I is the set of all items and S_1 a subset of the set of items ($S_1 \subseteq I$), a transaction $t_i \in T$, where T denotes the set of all transactions, is said to contain itemset S_1 if S_1 is a subset of t_i. The support count, $\sigma(S_1)$, for an itemset S_1 is defined as $\sigma(S_1) = |\{t_i \mid S_1 \subseteq t_i, t_i \in T\}|$, where $|\cdot|$ stands for the cardinality, that is, the number of elements in a set.

An association rule is then an implication expression of the form $S_1 \rightarrow S_2$, where $S_1, S_2 \subseteq I$ and $S_1 \cap S_2 = \emptyset$. The support, $s(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{|T|}$, determines how often a rule is applicable to a given data set. The confidence, $c(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{\sigma(S_1)}$, determines how frequently items in S_2 appear in the transactions that contain S_1.

5 OVERVIEW OF THE INCLUDED PUBLICATIONS

The goal of the thesis is twofold: to provide contributions to educational domain knowledge discovery and methodology. This chapter provides an overview of the 10 included publication that contain the contributions. First, it is explained how the single publications are connected, build on each other, and belong together (Section 5.1). Second, the 10 studies are discussed separately in a more detailed way (Section 5.2–Section 5.11). Third, the main results from both perspectives, that is, knowledge discovery and methodology, are tabularly summarized (Section 5.12).

5.1 Coherence and cohesion of the included publications

All the papers in this thesis contribute to knowledge discovery from sparse data from Finnish educational institutions or related to the management of a national educational system. The first article in this thesis, **PI**, introduces an analysis framework for sparse educational data. Through triangulation of supervised and unsupervised methods and the introduction of a ranking system, it is demonstrated that general study capabilities predict the study success of DMIT students better than specific IT skills. From the methodological vantage point, robust clustering that uses the spatial median as location estimate to assess the center of a set of points (see Section 4.2)—in particular, the initialization when there are missing data—is further advanced. As explained in Section 4.2, the robust clustering method employed in this article has the same algorithmic skeleton as the popular `k-means` algorithm, which is known for its initialization challenges.

Publication **PII** introduces the PISA data and provides an overview of publications based on PISA. It identifies the research gap (because of the technical complexities within the different representations of LSEA data and the lack of methods that allow advanced analysis of these large data sets, there is little research activity on the secondary analysis of these data), and describes the characteristics of PISA and general LSEA data that have to be taken into account when developing methods to work with these data. Moreover, **PII** includes a case study in which all the PISA 2012 countries are clustered hierarchically by taking for all countries the mean as input for each variable. Through statistical testing on different levels (by employing the triangulated approach proposed in **PI**), Finland's position in the international educational perspective is emphasized. It is concluded that Finland's comprehensive school system is able to cope with the challenges of negative attitudes toward mathematics, low work ethic, and little study time outside school by promoting student collaboration, humility, and equity (see Section 1.2).

Publication **PIII** is a knowledge discovery study that focuses on the Finnish subset of the 2012 PISA data. In comparison to **PII**, all students are treated as single entities; that is, no aggregating or averaging is employed. To deal with the size and the sparsity, the robust partitional clustering method described in **PI** is utilized. As explained in Section 4.3.1, hierarchical clustering is not scalable to a large number of observations, whereas partitional clustering algorithms are very feasible even for big data, and the spatial median with the available data strategy is used to handle the missing data. Moreover, the analysis framework introduced in **PI** (that is, triangulation of different analysis methods) is applied in **PIII**. The clustering of the Finnish sample yielded two obvious—one with high- and one with low-performing students—and two interesting—both with medium performing students—clusters. Association rule mining for the interesting clusters revealed very gender-specific characteristics for the two medium performing cluster: Average performing girls had very high attitudes toward school and learning in general, but no intentions to utilize mathematics later in life, while the average performing boys showed exactly opposite characteristics: They had the greatest intentions to pursue a mathematics-related career, but they did not like school in general.

Publication **PIV** extends **PIII** from a sample to a population level. To enable working with all the specific characteristics of PISA data discussed in Section 3.3 (i.e., the sizes, sparsity, and weights), it introduces the weights to the robust clustering algorithm for sparse educational data. The weighted robust clustering algorithm is based on the `k-spatial-medians` clustering employed in **PI** and **PIII** but can cluster a sample on a population level by initializing and updating the clusters with regard to the weights, that is, the importance of the single observations. This weighted robust algorithm is utilized for the Finnish subset of the PISA 2012 data in the same article and for the global PISA 2012 data in **PV**. From the domain knowledge discovery point of view, the findings in **PIV** were almost the same as in **PIII**, that is, the gender differences in the average performing clusters were present too.

Article **PII** uses only the aggregated global PISA data to emphasize Finland's position in the international educational perspective, but article **PV**, **PVI**, and **PVII** utilize all single observations of the global PISA 2012 data. In **PV**, the whole algorithm from **PIV** is applied in a hierarchical fashion for the entire PISA 2012 data. First, the PISA scale indices were used as input for the weighted clustering algorithm from **PIV**, which yielded two global clusters. Then, these global clusters were used as input for the same algorithm and so on, so that in the end, a cluster tree was created with three different levels of abstraction. This tree shows the hierarchies in the global PISA data. When a clustering result, such as the one in **PV**, exists, it still has to be evaluated for the quantitative educational knowledge discovery (see Section 2.1). **PVI** presents two novel methods that extend the Kruskal-Wallis test statistics for real-valued weights to evaluate such weighted clustering results—again by employing the triangulated approach proposed in **PI**. They thus advance the automatic educational knowledge discovery. These proposed methods are utilized to automatically evaluate the clustering result from **PV** and to rank the features and meta data according to their significance for the final cluster creation. It is found that the students' economical, social, and cultural status is the feature that explains the clusters of the global PISA hierarchical tree the most.

In article **PVII**, a PCA version is introduced that can handle sparse data. Similarly, as in the robust clustering method of articles **PI–PV**, which (as explained above) is algorithmically based on the traditional `k-means`, the algorithmic skeleton of the introduced robust PCA is based on the traditional PCA (see Section 4.3.2). However, instead of utilizing the sample covariance matrix that is based on the sample mean—and again similarly as in the robust clustering method, which uses the spatial median instead of the sample mean as the location estimate—the robust covariance matrix corresponding to the spatial median is employed. Triangulation (see again **PI**) of different robust PCA versions and comparison to the classical PCA showed that the robust approach is especially preferable when dealing with a high percentage of contaminated data. The robust PCA is also used for the PISA data, which again emphasized the importance of the students' economical, social, and cultural status.

Articles **PII–PVII** focus on the sparsity in the contextual PISA data, while article **PVIII** addresses the sparsity in the cognitive PISA data. As such, an algorithm was proposed to assign each student to a proficiency bin based on his or her raw test scores. Moreover, different classifiers were compared in their prediction accuracy by using only the students' raw answers (i.e., none of the already preprocessed and transformed variables, such as the PISA scale indices utilized in **PII–PVII**) to the background questionnaire as features. A sum of rankings (again as proposed in **PI**) of different feature selection algorithms on the raw questionnaire data showed that the self-evaluation on getting good grades in mathematics predicts the Finnish students' mathematics performance the most.

Article **PIX** is a follow-up study on **PI**. The newer study records from the same source as in **PI** are used. The robust clustering method with the initialization developed in **PIV** and **PV** is employed to find students with similar study paths. Together with a proposed architecture of an academic advising system, it is argued that these can be used for more automated and evidence-based decision making in an educational institution.

The robust PCA method introduced in **PVII** and the analysis framework for sparse data introduced in **PI** were used in **PX**. Through the triangulated analysis proposed in **PI**, publication **PX** demonstrates that most of the expert-based rankings in the Finnish publication channel evaluation system (see Section 3.4) can be predicted and explained using automatically constructed data mining and machine learning reference models. Finally, it is shown that those publication channels, for which the Finnish expert-based rank is higher than the estimated one, are mainly characterized by higher publication activity in combination with or solely by the recent upgrade of the rank. This leads to the assumption that a machine based ranking may be even more accurate and objective compared to human decisions.

5.2 Article PI: Analysing Student Performance Using Sparse Data of Core Bachelor Courses

This article was published in 2015 in the Journal of Educational Data Mining, Vol. 7.1, pages 3–32.

Objectives

The objective of this article was to identify the characteristics and structure of educational data mining studies and to establish a general educational data science (see Section 2.2) framework. Moreover, this article includes a case-study of assessing the study record data of bachelor students at DMIT (see Section 3.5) with the established framework.

Types of missing data and strategy to deal with them

The missing data values in the matrix that models the DMIT students' mandatory bachelor courses grades are MAR. This means that the missing values are related to particular variables (some courses that are usually taken later in the program are completed by fewer students; see Figure 2 in **PI**), but not missing because of the values (grades) that could be observed if a particular course is passed. Thus, the grade of the missing course is related to time, that is, the number of semesters the student has studied already. To deal with the sparsity, the correlation analysis is based on the available data only. Clustering was performed with the robust clustering algorithm for missing data (see Section 4.3.1). For the MLP prediction, the data was hot-deck imputed using the robust clustering result for a larger number of clusters, which was assessed to be a good number of clusters with the elbow curve and that still had complete centroids.

Contributions and results

A representative set of existing educational data mining studies was summarized according to a) their data and its environment, b) the goal of the study, c) the educational data mining category and the used methods, and d) the knowledge obtained. According to this summary, only methods belonging to one of the classes in the taxonomy by Baker (2010) (as listed in Section 2.2) are usually applied to address a particular educational data mining problem represented through data.[1] It was proposed that in comparison to the current state of the art, a particular educational data mining problem should be addressed through various approaches and assessed as a whole to increase both the technical soundness of the procedures and the overall reliability of the concluded results. The approach based on multiphase methodological triangulation (Denzin, 1970; Bryman, 2003) was introduced: different phases of the overall educational knowledge discovery process with both within methods and between methods were varied by their metaparametrization and then combined and assessed as a whole through a ranking system. More precisely, *prediction* using MLP (see Section 4.4), *clustering* using statistically robust procedure based on `k-spatial-medians` (see Section 4.3.1), *relationship mining* using two variants of correlation analysis and assessment of MLP's analytic feature saliency, and *discovery with models* by triangulating and ranking the results of individual approaches were employed. The *distillation of results for human judgment* was that the quality and efficiency of the bachelor studies at DMIT are very much determined by the first introductory courses. Based on the triangulated analysis, it was concluded that general study and learning capabilities predict the students' success better than specific IT skills learned as part of the core studies. Methodologically, how to cope with the non-structured sparsity pattern (i.e., the set of missing values) in data with both descriptive and predictive methods was shown. Moreover, the initialization for the robust clustering algorithm for sparse data was advanced to ensure that the resulting centroids needed for the educational knowledge discovery are complete. In summary, the small complete data set (students who completed all courses) was used to determine the best centroids and best number of clusters, and the resulting centroids were iteratively used again to initialize the next larger subset (students who completed all but one course), which again were used as input for the succeeding larger subset.

Author's contributions

The author of this thesis is the main and corresponding author of this journal publication. She preprocessed the data, carried out the first two parts of the triangulated data analysis, and provided the through hot-deck imputation completed data for the third part of the analysis. Moreover, the author of this thesis produced all tables and figures of this article (except those in Section 5), interpreted the results, and wrote the majority of the paper, with the exception of Section 5.

[1] Moreover, in the prediction category, different classifiers are often compared, whereas for the other categories, just one method from existing data mining tools or libraries is usually utilized.

5.3 Article PII: Knowledge Discovery from the Programme for International Student Assessment

This article was published as the eighth chapter in the 2017 Springer book *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art*, pages 229–267.

Objectives

The objective of this article was to provide the general background for the empirical PISA part of this thesis: first, a comprehensive overview of PISA data and their characteristics and, second, a coverage of related work concerning scientific papers about PISA, research related to the high performance of Finland in PISA, and educational clustering studies. Moreover, the goal was to cluster all the PISA countries to identify Finland's position within the international context and to identify the strengths and shortcomings of the Finnish in comparison to the global learning environment.

Types of missing data and strategy to deal with them

Most of the missing data are MCAR by design. The data was clustered hierarchically by taking the country mean for each variable.

Contributions and results

It was found that most of the related work concerned with analyzing PISA data are national and international reports, but only a few studies have been published in scientific publication channels where the articles have to endure the anonymous peer-review process. Moreover, those scientific PISA data articles that were identified mostly followed the traditional hypothesis testing research approach (see Section 2.1), and many did not take the special characteristics of PISA data (described in Section 3.3) into account; for example, they analyzed only the sample by ignoring the weights, lost a large subset of data by discarding observations with missing data, and utilized only a small subset of the PISA data by focusing on only a few countries. From the literature review on educational clustering studies, it was concluded that most of these studies are based on hierarchical or representative-based (usually, k-means or expectation-maximization) methods. From the case study, it was found that requiring only minimal effort (latest school start from all PISA countries, almost no homework, and very manageable amount of hours in school) from the Finnish students and the equal treatment of every individual lead to one of the least-motivated student cohorts of all the PISA countries. Students have no ambition to excel, a low work ethic, and only very marginal motivation to pursue a mathematical-related career. The further dropping of Finland in the international mathematics performance ranking of the succeeding PISA assessment (that de facto occurred; see Figure 1 in the introduction of this thesis) was predicted. However, it is known that in PISA, some attitude variables are positively correlated with achievement within a country,[2] but negatively correlated at the country level (Kyllonen and Bertling, 2014). This might explain parts of the results. As also explained in this article, Finnish citizens are rather modest about their own achievements, and they place great emphasis on equity and equality. The most important driving factors in the life of this highly feminine country are to live a good life and to care for others rather than to focus on one's own success and desire to be the best. Therefore, they might report some of the attitude variables less enthusiastically than students from other countries.

Author's contributions

The author of this thesis is the main and corresponding author of this book chapter publication. She conducted the literature review on research related to learning analytics and the high PISA results of Finland; preprocessed the data; designed, implemented, and carried out the data analysis and statistical tests; produced all tables and figures (except Figure 18.12); interpreted the results; and wrote the majority of the article. The literature review on educational clustering studies (Section 8.2.2) was mainly written by the second author, but the author of this thesis collected the described work from the main publishing forums for LA studies. Moreover, the author of this thesis presented preliminary results of this study at the Annual Computer Science Event 2015 in Jyväskylä, Finland.

5.4 Article PIII: Discovering Gender-Specific Knowledge from Finnish Basic Education Using PISA Scale Indices

This article was published in the full paper proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pages 60–68.

Objectives

Proficiency in mathematics strongly predicts admission to post-secondary education and expected future earnings of adolescents (p.252 OECD, 2014a). According to the official reports by the OECD, Finland was one of the few countries in the 2012 PISA assessment in which girls performed slightly better in mathematics than boys, and the Global Gender Gap Report by the World Economic Forum (2016) ranks Finland worldwide second (after Iceland) in gender equity. However, in Finland mathematics-related jobs are also dominated by men. The purpose of this study was to refine the analysis of this observation by leveraging data mining techniques for the educational domain using the PISA 2012 questionnaire

[2] See, for example, **PIII** where it was found that for Finnish students, the self-concept in mathematics is highly and positively correlated with the plausible values in mathematics.

scale indices that are known to affect proficiency in mathematics.

Types of missing data and strategy to deal with them

The missing data in the utilized PISA questionnaire scale indices are MCAR by design, as each student is, as explained in Section 3.3, administered one out of three different background questionnaires from which these indices are created, as part of the arrangement of the assessment. To deal with the sparsity, the robust `k-spatial-medians` was employed.

Contributions and results

It was found that the Finnish students sample divides into four clusters. Two of these clusters unambiguously could be explained by performance: One cluster consisted of the very high-performing Finnish students and the other cluster consisted of the low-performing students. The remaining two cluster were composed of medium-performing students. With association rule mining, it was revealed that the students in those two average performing cluster groups have very gender-specific attitudes: Girls had the highest attitudes toward school and learning in general, but no intentions to pursue a mathematics-related career. Boys were not interested in school and learning at all, but had the highest expectations of leveraging mathematics in their future. Methodologically, the robust cluster initialization for sparse data was further advanced by (comparable to the procedure proposed in **PI**) using only the complete data first to assess with a cluster index the best number of clusters, *K*, through multiple repetitions and then using the centroid from the complete data and the determined best *K* for the initialization of the full data.

Author's contributions

The author of this thesis is the main author of this publication. She preprocessed the data, carried out the data analysis, produced all tables and figures, interpreted the results, and wrote the majority of the paper. Moreover, the author of this thesis presented the paper at the 7th International Conference of Educational Data Mining in London, UK.

5.5 Article PIV: Weighted Clustering of Sparse Educational Data

This article was published in the proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015), pages 337–342.

Objectives

The objective of this article was to enhance the study in **PIII** by clustering not only the PISA sample but the whole Finnish population of 15-year-old students. To establish this, the sampled students must be clustered not as single entities but with their weights, that is, the numbers that depict for each of these particular students how many students in the population he or she represent (as explained in Section 3.3). Weighted clustering is not very widely addressed (Ackerman et al., 2012) and, for example, not part of standard data mining software. Thus, the ambition of this article was to propose an efficient version of the robust clustering algorithm for sparse data (i.e., the `k-spatial-medians` algorithm from the previous studies) that takes the weights, aligning a sample with the corresponding population, into account.

Types of missing data and strategy to deal with them

The data used in this article were the same as in **PIII**, which means that the missing data are MCAR. The weighted version of the further developed robust clustering `k-spatial-medians` algorithm was used to deal with the sparse and weighted data.

Contributions and results

An efficient version of a robust weighted clustering algorithm was introduced. This algorithm takes the weights into account in all steps of the algorithm where they are needed. Similarly to the initialization approach in **PIII**, cluster indices and multiple runs of the algorithm with complete data were utilized to find the best initialization for the full data. However, in this article, more cluster indices were used for comparison, and the cluster indices were modified to work with the weights. After the initialization, the weights had to be taken into account only in the second iterative step: Each observation is assigned to its closest centroid, but the update of the centroid should be more in the direction of the more important observations, that is, those observations with larger weight. From the domain point of view, the results of this study where the Finnish population of 15-year-old students was clustered were very similar to the results where the Finnish PISA sample (**PIII**) was clustered.

Author's contributions

The author of this thesis is the main author of this publication. She preprocessed the data, carried out the data analysis, produced all tables and figures, interpreted the results, and wrote the majority of the paper.

5.6 Article PV: Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data.

This article was published in the full paper proceedings of the 8th International Conference on Educational Data Mining (EDM 2015), pages 156–163.

Objectives

As explained in Section 1.2, a relationship between culture and attitudes exists. In particular, it has been argued that culture affects people's goals and their actions to reach these goals (Hitlin and Piliavin, 2004). The research question of this article was as follows: If all 24 million students in the PISA data were clustered as single entities with their characteristics and attitudes, but without using the country information in the clustering algorithm, could the resulting clusters be explained by their country? To realize such a clustering procedure that can cope with the PISA data characteristics (i.e., the sparsity, weights, and sizes), the objective of this article was to carry out the weighted robust partitional clustering algorithm from **PIV** hierarchically for the entire PISA data.

Types of missing data and strategy to deal with them

The variables used in this article were the same as in **PIII** and **PIV**, which means that the missing data are MCAR. However, in this study, not only the Finnish subset was utilized but also all observations in the PISA data of all participating economies and countries with their weights. Because of the design of the assessment, the percentage of missing data for the global level, that is, the entire PISA data, is almost the same as the percentage of missing data on the country level. The weighted version of the robust clustering `k-spatial-medians` algorithm from **PIV** was used to deal with the sparse data.

Contributions and results

As pointed out in Section 4.3.1, hierarchical clustering is only feasible for very small data sets. In this article, partitional clustering (i.e., the `k-spatial-medians` algorithm), which is very scalable and feasible for large data sets, was applied in a hierarchical fashion so that a hierarchical tree of clusters could be established without the demand to employ an expensive hierarchical clustering algorithm. The initialization and determination of the number of clusters for each subcluster of the hierarchical tree was performed as in **PIV**. From the domain level point of view, it was found that performance in the PISA tests can explain the discovered clusters, but the actual country data contributes to the cluster membership information only to a marginal extent.

Author's contributions

The author of this thesis is the main author of this publication. She preprocessed the data, carried out the meta data analysis, produced all figures and tables, interpreted the results, and wrote the majority of the paper. The author of this thesis also presented the paper at the 8th International Conference of Educational Data Mining in Madrid, Spain.

5.7 Article PVI: Feature Ranking of Large, Robust, and Weighted Clustering Result

This article was published in the full paper proceedings of the 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2017), Springer International Publishing.

Objectives

To complete the (educational) knowledge discovery process, as described in Section 2.1 of this thesis, the numerical results from the data mining step have to be interpreted. When the data are clustered, that means that the data (originally clustered data and optionally metadata) with the found cluster labels can be evaluated for the interpretation. To facilitate such an interpretation, the Kruskal-Wallis test statistics can be utilized to establish a ranking of variables and, therefore, to discover which variable affected the cluster creation the most. However, when clustered data represents a sample from a population with known sample-to-population alignment weights (as in **PIV** and **PV**), both the clustering and the evaluation techniques need to take this into account. The objective of this article was to introduce the weights to the Kruskal-Wallis test statistic to advance the automatic knowledge discovery from a population-level clustering result. This is a difficult problem in statistics since the Kruskal-Wallis test depends on data ranking.

Types of missing data and strategy to deal with them

The missing data in PISA data are MCAR by design. Only the existing values are used to compute the weighted Kruskal-Wallis test statistics.

Contributions and results

Two different approaches were suggested that can rank the variables of a weighted clustering result by generalizing the Kruskal-Wallis test statistic from the sample to population level. The first approach extends the integer value weights approach suggested by Tölgyesi et al. (2014), which involves copying each observation as many times as the integer weights suggest, for real valued weights by using the classical bootstrapping (Efron, 1979). The second suggested approach is based on a novel heuristic formula derived in the article. To test the approaches, the clustering result from **PV** was utilized. Both the input data (data that was clustered in **PV**) and the metadata, that is, all the variables from the PISA information and communication technology (ICT) questionnaire (OECD, 2015), were used to compare the data distributions in the existing clusters with the proposed approaches. As hypothesized in **PV**, it was found that the students' economic, social, and cultural status (ESCS) is the most important variable determining the different clusters. The plausible values (PVs) were found to be the most important variables from the metadata. This result is reasonable because the input/clustered variables are, as explained in Section 3.3.1 of this thesis, part of the posterior model from which the PVs were sampled. Moreover, the approaches were compared using the labeled Iris data from the UCI machine-learning repository (Merz and Murphy, 1998), and this methodological triangulation also showed that the results were very consistent between the different approaches. Finally, using the analytical formula for quick evaluation was recommended. The bootstrap approach (which is better aligned to the existing literature) was recommended for automatic clustering result rankings and for finalizing the educational knowledge discovery process.

Author's contributions

The author of this thesis is the first author of this article. She preprocessed the data, described the PISA data, interpreted the results, and wrote the corresponding sections of the article. Moreover, the author of this thesis presented the paper at the 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining in Jeju (South Korea).

5.8 Article PVII: Robust Principal Component Analysis of Data with Missing Values

This article was published in the full paper proceedings of the 11th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2015), pages 140–154.

Objectives

Although PCA is one of the most popular methods in data mining and machine learning, the use of PCA for sparse data with missing values seems not to be a widely addressed topic. The objective of this article was to propose multiple robust PCA approaches for data with missing values and to estimate the relative importance of the principal components to explain the data variability.

Types of missing data and strategy to deal with them

The missing data were assumed to be missing completely at random. For the carefully designed test data sets, a MCAR sparsity pattern was introduced artificially. Missing data in PISA data are MCAR by design. The proposed PCA approach uses the available data strategy and spatial median location estimate to calculate a more robust covariance matrix, which is then used to compute the principal components.

Contributions and results

A novel PCA approach was introduced, which uses the robust covariance matrix Σ_R that corresponds to the spatial median (the location estimate explained in Section 4.2) instead of the sample covariance matrix that is based on the mean. Since the robust covariance matrix Σ_R is based on first-order approximation, the eigenvalues readily correspond to the geometric variability represented by the standard deviation. As argued in Section 4.3.2, the eigenvalues of the standard covariance matrix Σ represent the *variances* along the new coordinate system, whereas to assess the true *variability* of data in one direction, the standard deviation should be used to determine the importance of a principal component. That means that in the standard PCA, the square roots should be taken to assess the relative importance of the components and to ensure that the robust and the classical PCA approaches are comparable to each other. Experiments were carried out that focused on carefully designed simulated tests where the ground truth was known and could be used to assess the accuracy of the results obtained with the robust in comparison to the traditional approach. Moreover, two modifications of the robust PCA were introduced based on the "almost complete data" concept (comparable to the initialization idea of the robust clustering in **PI**, **PIII**, **PIV**, and **PV**) to prevent the underestimation of the amount of variability of data and/or the main directions of variability due to sparse data vectors. The first modification uses the 10% and 90% percentiles of the projections along the new coordinate system, and the second modification uses only those observations of which one variable at most is missing to estimate the robust covariance matrix Σ_R. The robust PCA approach yielded good results (especially with the first modification that used percentiles to estimate the relative importances of the principal components) when compared with the true variability of the data, and the estimated directions remained stable even with a large amount of missing data, whereas the classical PCA was more prone to nongaussian errors in the data. Moreover, it was shown that even if there are missing values in the original data, the resulting new data vectors become complete when the robust PCA is employed, while they have missing values when the traditional PCA is employed. The introduced robust PCA approach was also applied to the same global PISA data set as in **PV**, where about 30% of the data were missing, and the data was projected to the first two principal components. This visualization showed that the students' economic, social, and cultural status divides the participating PISA 2012 students the most, which confirms existing the information. For example, as pointed out by the OECD "PISA 2012 data shows that the economic, social and cultural status of students explains 24% of the mathematics performance variation among all PISA countries and even 46% of the variation among OECD countries" (OECD, 2014a, page 36).

Author's contributions

The author of this thesis carried out the experiments, produced all figures and tables, interpreted the results, and wrote the corresponding sections of the article. Moreover, the author of this thesis presented the paper at the 11th International Conference on Machine Learning and Data Mining in Hamburg, Germany.

5.9 Article PVIII: Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach

This article was published at the website of the Machine Learning for Digital Education and Assessment Systems workshop of the 33rd International Conference on Machine Learning (ICML 2016).

Objectives

PISA reports student proficiencies only in the form of plausible values (as explained in Section 3.3.1). Plausible values have shown to be a reliable estimate for proficiencies of populations and are used not only in PISA but also in all major large-scale educational assessment studies (listed in Section 3.2). However, a more comprehensive study of PISA data sets by deploying machine learning algorithms may provide a better understanding of the underlying factors affecting student performance and thus yield to better and more interpretable predictive models. Although the Rasch model with the plausible value approach currently used in PISA has been criticized, for example, by Kreiner and Christensen (2014), no real alternative to analyze the sparse cognitive performance data of these assessments has been suggested (Adams, 2011). The objective of this article was to perform a supervised approach for PISA data to predict students' performance in mathematics without using any of the derived variables in PISA but only the raw really observed data. This is a challenging task because of the high sparsity in the scored cognitive response data.

Types of missing data and strategy to deal with them

The missing data of the scored items in the PISA cognitive test data are missing by design. This data set provides for each student the information about whether an item was administered to this student and, if so, how many points the student received for this item. However, all the students that were administered the same cognitive test have a complete matrix for that set of items included in that specific test. To deal with the sparsity, first, a seven-dimensional matrix was constructed that assigned each item to a bin depending on how many students (out of all students from the same test) were able to solve this item correctly. Then, each student was assigned to a bin depending on whether he or she mastered the corresponding difficulty level of a bin.

Contributions and results

A technique was presented to learn directly from LSEA data by deploying a combination of both unsupervised and supervised learning feature selection algorithms to predict student performance on mathematics scores. The technique learns the difficulty level of different mathematic task items and predicts whether a student with a particular background profile will be successful in answering correctly. For this, all raw answers directly related to the performance in mathematics were utilized from the PISA 2012 students' background questionnaire. Moreover, since correct answers for easier questions are predictive for harder ones, the information about whether the student mastered the previous difficulty level(s) was iteratively added to the original feature set of raw mathematics background questionnaire answers. Preliminary results of different supervised data mining and machine learning models were compared in their prediction accuracy for the created labels. The algorithm was tested first for the Finnish students only and then for all participating students. For the Finnish students, the variable assessing the students' self-concept of getting good grades in mathematics was the most important variable for predicting the performance in mathematics, while the students' subjective norm that their parents like mathematics was not important at all.

Author's contributions

The author of this thesis is the main author of this publication. She implemented the data analysis, produced all figures and tables, interpreted the results, and wrote the majority of the paper. Moreover, the author of this thesis presented the paper at the Machine Learning for Digital Education and Assessment Systems workshop of the 33rd International Conference on Machine Learning in New York, USA.

5.10 Article PIX: Supporting Institutional Awareness and Academic Advising Using Clustered Study Profiles

This article was published in the full paper proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017). Moreover, an extended version of this paper has been selected to be included in the Communications in Computer and Information Science series published by Springer.

Objectives

Academic advising (i.e., the process of helping individual students with developing educational plans and guidelines that support their academic career and personal goals) is associated with a high work load for the academic advisor and therefore costs for higher education institutions that employ academic advisors. The objective of this article was to propose a system that performs academic advising in a more automated matter.

Types of missing data and strategy to deal with them

The missing data of the study records that are used to demonstrate the operation of the proposed system are MAR because the missingness is related to time (see **PI**). The robust clustering method with the available data strategy and initialization for sparse data (see **PIII**-**PV**) was used to handle the sparse data.

Contributions and results

A type of automated academic advising was suggested that is based on real study records. This system takes the sparse data matrix created from the study log of passed courses (see Section 3.5) as input and uses robust `k-spatial-medians` clustering with initialization for sparse data to identify a set of actual study path profiles. Such profiles identify groups of students with similar progress of studies, whose analysis and interpretation can be used for better institutional awareness and to support evidence-based academic advising.

Author’s contributions

The author of this thesis described the data and the method of the proposed system. Moreover, the author of this thesis revised the paper and explained the advantages of using clustering. The data analysis with the initialization for sparse data, which was performed by the third author, corresponds to the approach evolved in **PI**, **PIII**, **PIV**, and **PV**.

5.11 Article PX: Expert-Based versus Citation-Based Ranking of Scholarly and Scientific Publication Channels

This article was published in 2016 in the Journal of Informetrics 10(3), pages 693–718.

Objectives

The objective of this article was to assess whether the expert-based rankings in the Finnish publication forum (see Section 3.4) can be constructed automatically through machine learning and data mining techniques. For this, all available variables from three databases (the Finnish databases containing the publication source information, the database containing the actual national publication activity information, and Thomson Reuters' Journal Citation Reports) were used that could affect the ranking of a publication channel (most importantly, the rank in other major citation databases, the age, the publication language, the type, and the rank in previous years). These variables are also the same that are provided to the expert panels to judge and rank the publication channels.

Types of missing data and strategy to deal with them

The missing data in the analyzed data from the three databases are NMAR; that is, the missing values depend on the value that would have been observed if the value had not been missing. The information related to whether a publication channel is indexed in the major citation databases was encoded and employed with association rule mining. The obtained rules showed that this information is actually an important predictor of the rank.

Contributions and results

It was demonstrated that most of the expert-based rankings can be predicted and explained using the automatically constructed reference models based on association rule mining, decision trees, and confusion matrices. Moreover, it was found that those publication channels, for which the Finnish expert-based rank is higher than the estimated one, are mainly characterized by higher publication activity or a recent upgrade of the rank. It was concluded that the large correspondences of the expert-based ranks with other information could allow researchers to partially automatize the manual ranking process or, at least, provide an accurate baseline for human decision making in the evaluation panels.

Author's contributions

The author of this thesis is the main and corresponding author of this journal publication. She conducted the literature review on research related to indicators that measure scientific output based on publications, preprocessed the data, implemented and carried out the data analysis, produced all figures and tables, interpreted the results, and wrote the majority of the article, except the literature review on performance-based funding systems in other countries (the first part of the introduction), which was mainly written by the second author.

5.12 Summary of contributions

This thesis is composed of the articles discussed above that contain the contributions. As emphasized before, the two main contributions of the publications are, first, adaption of computational methods to data with special characteristics, especially LSEA data and, second, applications of existing and further developed methods for sparse educational data sets, especially the data from the 2012 PISA assessment. Table 2 summarizes the main contributions from these two perspectives for each publication and acknowledges the data that were utilized. As explained above, the publications comprising this thesis utilize three different data sources, that is, the PISA data (Section 3.3), the data containing information about the publication sources and publication activities of Finnish researchers (Section 3.4), and the DMIT study records (Section 3.5).

TABLE 2 Results.

Article	**Data**	**Methodological Orientation**	**Knowledge Discovery**
PI	DMIT students study record log data	Analysis framework for sparse educational data, correlation analysis, robust clustering, MLP neural network prediction	General study capabilities are more important for study success at DMIT than specific IT skills.
PII	Global PISA contextual student data	Combination of data mining method (hierarchical clustering of weighted means of each country) with statistical testing of the results	In comparison with their international peers, Finnish students can be characterized by their high ESCS but very low work ethics and motivations to study.

Continued on next page

Table 2 – continued from previous page

Article	Data	Methodological Orientation	Knowledge Discovery
PIII	Finnish subset of PISA 2012 contextual student data (scale indices associated with mathematics performance)	Use of the analysis framework from **PI**; sparse data analysis; triangulation of different analysis methods: correlation analysis, robust clustering, association rule mining	Average performing girls and boys have very gender-specific attitudes concerning mathematics and their future career.
PIV	Finnish subset of PISA 2012 contextual student data (same as in **PIII**)	Incorporating the weights into the robust clustering algorithm from **PI** and **PIII**	Similar clusters as in **PIII**
PV	Global PISA contextual student data (same variables as in **PIII** and **PIV**)	Hierarchical operation of the weighted robust clustering algorithm from **PIV**	Performance in the PISA test explains global PISA clusters more than country of the student.
PVI	Global PISA 2012 contextual student data	Introduction of weights to the Kruskal-Wallis test statistics to allow variable rankings for automatic knowledge discovery and interpretation for clustering results on the population level	ESCS most important input variable for determining cluster membership in **PIV**, PVs most important from the meta-variables.
PVII	Global PISA 2012 contextual student data (same variables as in **PIII**-**PV**)	Novel robust PCA version for data with missing values	Confirmation of existing knowledge: ESCS is the most separating variable in entire PISA 2012 data.

Continued on next page

Table 2 – continued from previous page

Article	Data	Methodological Orientation	Knowledge Discovery
PVIII	Global PISA 2012 cognitive data (for establishing the labels) and raw contextual data (as predictors)	Establishing performance labels for students based on the sparse cognitive PISA data, using different classifiers and raw questionnaire data to predict performance	The self-concept on getting good grades is the raw mathematics variable that predicts mathematics performance the best.
PIX	DMIT students study record log data (follow-up data from **PI**)	Sparse data analysis; use of the robust clustering with the initialization developed in **PIII**-**PV**	Similar student profiles that can be used to automate academic advising.
PX	Database containing the Finnish expert-based rankings of publication sources and database containing the Finnish national publication activity information.	Sparse data analysis; use of the analysis framework from **PI**; triangulation of different analysis methods: association rule mining, decision trees, and confusion matrices; use of the sparse PCA approach from **PVII**	Most of the expert-based rankings can be predicted and explained using automatically constructed reference models. Publication channels, for which the Finnish expert-based rank is higher than the estimated one, are mainly characterized by higher publication activity or recent upgrade of the rank.

6 DISCUSSION AND CONCLUSIONS

The high results in the firsts PISA assessments have made Finland's educational system internationally famous, and since then, this system has been under active study but mainly by educational scholars using traditional analysis techniques and manually collected data. The work of this thesis complemented the existing research by analyzing and discovering knowledge from the Finnish educational system in the large—that is, basic education (PISA, see Section 3.3), higher education (university, see Section 3.5), and the resource allocation for higher education (Jufo, see Section 3.4)—and by leveraging and further evolving data mining methods for the educational domain (as highlighted in Figure 5). These methods incorporate more multivariate techniques than classical educational research and are suitable for data that are large, both in terms of observations and dimensions.

To answer the research questions posed in Section 1.3, this thesis can be concluded from three different point of views: the context of automation (**RQ1**), the methodological development (**RQ2**), and the educational knowledge discovery perspective (**RQ3**).

RQ1 - Automation

From the context of automation, the educational knowledge discovery process is more mechanized than traditional educational research. Instead of manually collecting specific data, the researcher may work directly with rich (openly) available data. Moreover, the traditional data analysis approach relies on hypothesis development and testing, whereas in data mining, the machine learns from the data and might be able to detect interesting patterns in them without the need of forming strong hypotheses first. This may lead to unexpected and novel findings, such as the one in **PIII**, but does not rule out hypothesis testing and confirmatory research to properly assess these findings. In fact, data mining is just a part of the whole educational knowledge discovery process.

As emphasized throughout the entire thesis, the educational knowledge discovery process is a stepwise procedure (see Figure 4) where human judgment is used to assess and decide (i) which particular target datasets are processed further, (ii) what the goal of the data mining is, (iii) what preprocessing and transformation methods are needed and used, (iv) what data mining algorithm(s) are utilized, (v) how the evaluation and interpretation is made, and (vi) what knowledge is concluded. That means that humans and machines work together in the educational knowledge discovery process, often employing the computer to discover and summarize information that assists in complex decision making (see, e.g., **PI**, **PII**, **PX**) and leaving the final contextual decisions to human domain experts (Merceron et al., 2016). Thus, human judgment is heavily involved here, but the automation is linked to what happens *inside* these steps.

In the publications included in this thesis, automation inside the educational knowledge discovery process steps has been promoted. For example, the robust clustering with the available data strategy (**PI**, **PIII**, **PIV**, **PV**, and **PIX**) avoids explicitly considering how to perform imputation because it does not need imputation. Furthermore, internal cluster validation indices (**PI**, **PII**, **PIII**, **PIV**, **PV**, and **PIX**) suggest how many profiles are hidden in the clustered dataset without human involvement. Moreover, feature importance measures allow automatic ranking and detection of the most important input features and metadata variables to ease up the interpretation of a clustering (**PVI** and **PI**) or prediction (**PVIII** and **PI**) result. Finally, on the organizational level, the integration of discovered rules and relationships into systems/reference models supported more automated utilization of the results (**PX** and **PIX**). In summary, the automation of the educational knowledge discovery process has been advocated, but without understating the crucial rule of the human domain experts.

RQ2 - Methodology

From the methodological point of view, nonstandard techniques based on the so-called robust statistics were utilized and further developed. Since all the data sets analyzed in this thesis had a severe sparsity pattern (**PI–PX**), it was argued that the existing unsupervised data mining methods yield more accurate and reliable results when the spatial median instead of the sample mean is used to estimate the center of a group of points (**PI**, **PIII**, **PIV**, **PV**, **PVI**, **PVII**, **PIX**, and **PX**). The sample mean, which is the default location estimate in some of the most applied unsupervised data mining techniques (i.e., k-means clustering and standard principal component analysis) is known to be extremely sensitive toward all kinds of outliers including missing values. On the contrary, the spatial median with a breakdown point of 0.5 still produces reliable results even when half of the data is contaminated. Furthermore, the available data strategy utilized within the unsupervised techniques of the publications in this thesis ensures that all of the existing observations are used, and thus, none of the possibly valuable information gets lost. In fact, in **PX**, it was shown that using the missingness information of certain variables in the constructed patterns and rules can reveal very important information in the knowledge discovery process.

The included publications especially focused on the special characteristics of LSEA data, and PISA data (**PII–PVIII**) represent the most prominent example—

both with regard to size and political acknowledgment. Besides the high sparsity, another distinguished feature of LSEA data sets is the real-valued weights that align the participating student sample with the whole student population. These weights were incorporated into the robust clustering algorithm (**PIV** and **PV**), its initialization (**PIV** and **PV**), the cluster indices determining the number of clusters (**PIV** and **PV**), and to the Kruskal-Walllis test statistics for ranking the variables used in clustering, therefore advancing the automatic educational knowledge discovery process by providing a machine-aided process to determine the variables most significant for a clustering result (**PVI**).

The importance of multiphase methodological triangulation, that is, to look at the same problem through different approaches, has been emphasized throughout this thesis and its included publications (especially in **PI**, **PVI**, and **PX**). As already argued by Gifi (1991) (see also Olsen, 2005a), if several techniques lead to the same conclusion, it is more likely that these reflect genuine and overarching aspects of the data and that the interpretation is not just an artifact of one particular technique used to analyze or inspect the data. In several publications of this thesis (**PI**, **PII**, **PIII**, **PIV**, **PVI**, **PVII**, **PVIII**, and **PX**), different between and/or within methodological approaches (see, e.g., Bryman, 2004b; Denzin, 1970) were used and compared to ensure that the concluded results were not an artifact of a particular approach and, thus, increased the technical soundness of the procedures and the overall reliability of the research outcomes. This kind of analysis and ranking frame is also used in novel technologies and intelligent systems, such as IBM's Watson (see, e.g. Gondek et al., 2012), to arrive at a final decision. Hence, to answer and conclude research question **RQ2** in one brief sentence, methods for analyzing LSEA data sets should be characterized by their ability to handle size, sparsity, and weights in data and should have proved to be reliable and consistent in comparison with other methods applied to the same problem.

RQ3 - Knowledge discovery

From the knowledge discovery perspective, the publications included in this thesis covered three different educational domains in Finland. The main domain was the Finnish comprehensive schools and their 15-year-old student cohort represented by PISA data (publications **PII–PVIII**). The second domain was the Finnish higher education exemplified through the study program at DMIT (publications **PI** and **PIX**). The third domain was the Finnish research setting and its performance-based funding system (publication **PX**).

The findings from the Finnish comprehensive school seem contradictory. The average decline of the mathematical achievement of the Finnish students in PISA 2012 to the assessment in 2003, where mathematics was the main assessment area the last time, is equivalent to the progress usually made in more than half a school year (Välijärvi et al., 2015). However, it is amazing that Finnish students still achieve high results in PISA,[1] especially because they are, in com-

[1] Although the average performance of Finnish students strongly declined, Finland is still among the highest-ranking PISA countries in the world (Välijärvi and Sulkunen, 2016).

parison with their international peers, characterized by an extremely low work ethic (**PII**); that is, they seem to see no reason why they should strive for excellence. The Finnish educational system appears to support and serve students with learning deficiencies well. This is particularly because special needs education is integrated in regular schools as much as possible and students of all levels study in the same environment for the first 10 years (see Section 1.2). However, it seems that less attention and emphasis are placed on high-achieving and gifted students (Tirri and Kuusisto, 2013). Treating everyone equally and requiring only the minimal effort does not produce the most ambitious and eager students (**PII**). Of all the PISA countries, Finland is, for example, at the bottom of percentage of students who have heard of or know the concept of complex numbers (p. 166, Figure I.3.14 OECD, 2014a). Maybe schools and education must become more challenging in Finland. More research is also required to determine the effects of more support for skilled students in Finland. It would be interesting to assess how much such settings would affect the overall PISA ranking results.

On one hand, more attention could be paid to talented students. On the other hand, it seems to be exactly the collaborative, trust-based, and less competition-oriented school system and Finnish culture that has contributed to the high average PISA results. From their economical, social, and cultural status (the single most important predictor of performance in PISA; see **PVI** and **PVII**), the Finnish students are comparable to their Western and Nordic neighbors (**PV**). However, the Nordic/Western student cohort tends to be more ambitious and motivated, while the Finnish students are distinguished by their more collaborative thinking and general humility (**PII**). As also pointed out in the literature review in **PII**, research has shown that Finnish citizens commonly place great emphasis on equity and equality and are very modest about their own achievements. This definitely benefits less-skilled students and, according to Välijärvi et al. (2007), virtually does not harm higher achieving students. However, it would be interesting to assess the effects in the Finnish learning environment if *efforts* and *hard work* were rewarded more.

Following the data mining definition by Hand et al. (2001) (and similarly the original knowledge discovery process definition by Fayyad et al., 1996b), that is, analyzing the data with the goal of finding novel, interesting, and useful patterns in them, research question **RQ3** also asked about the usefulness of the discovered knowledge. In the end, this usefulness might be a very subjective measure that must be evaluated by the respective domain experts. However, for example, the automatic construction of rules for the ranks in the Finnish publication forum could be of direct practical value. As explained in Section 3.4, human decision making is associated with high costs for the Finnish government. Using the automatic constructed ranks from **PX** would save money and man-hours in the research organisation setting and bring more objectivity into the rankings. Similarly, the automated academic advising system from **PIX** could lead to economization of human study guide and recommendation efforts and increased awareness of the different students and their personal study paths in higher education. Nevertheless, the more work is allocated to a machine, the more important it becomes to define and implement ethical boundaries and rules determining, for example, what kind of information can be used in such systems while ensuring the protection of individual rights and their data (Hildebrandt, 2017).

The discovery that boys and girls who are attending Finnish basic education classes have different goals for their future (**PIII**) may not be so obvious in its value for immediate decision making. Nonetheless, it is interesting that on one hand, Finland seems to be one of the most developed countries in terms of equal rights for men and women. On the other hand, traditional values and views seem to be still deeply anchored in the population. Thus, this finding should be studied more. In particular, it would be interesting if further studies would examine at which age this gender-specific goals and associated future career wishes appear.

Research in gender and equity has shown that girls are more likely to chose a mathematical/programming related career if they believe they will be successful in it. However, they often assume that boys are better in these areas, and thus, they do not consider such a career in the first place (e.g., Denner, 2007). Hence, it is important to change the girls' own stereotypical behaviors. The finding that general study capabilities at DMIT are more important than mathematical and programming skills (**PI**) might serve as a motivator for girls to pursue and persist in attaining a career in this field. In PISA 2012, the gender difference in the mathematics performance of Finnish students was three score points in favor of girls (**PIII**). In PISA 2015, this difference has widened to eight score points (OECD, 2017a). Encouraging girls to consider a career in mathematics-related fields, and therefore to increase their interest in and excitement to exert themselves for it, may result in even higher average girls' performance and thus increase Finland's overall place in the international ranking of forthcoming LSEA studies.

Limitations and Future work

There are some limitations to the findings presented in this thesis and a number of interesting possibilities for future work that can extend the research started in here.

One obvious limitation of this thesis is that the discovered knowledge is restricted to those methods that were used. The methodological landscape of data mining and machine learning is immense, and thus, more methods could be tested and adapted to the specific requirements of LSEA data. In particular, future work should be undertaken to try to improve techniques concerning sparsity and weights. These techniques could be used to test and verify the discovered knowledge or even lead to more novel information.

Another limitation arises from the educational knowledge discovery point of view discussed in Chapter 2. The empirical part of this work did not include any data collection but focused on existing data. Most of the analyzed data sets, especially PISA, are very much processed and if one looks at the big picture, there could be more work regarding the actual raw data. The study in **PVIII** already presented an approach for utilizing the raw PISA data and this will be tested more in the future. Moreover, it is currently under investigation how the PISA log file data that contain even more raw data (and that have been made available for the electronic PISA tests, see **PII**) can be leveraged for the educational knowledge discovery process.

A further natural progression of this work is to analyze data from forthcoming PISA cycles and other LSEA studies with the proposed methods. Further work related to Finland's drop in the PISA mathematics ranking was proposed along with concluding **RQ3**. As discussed there, it would be particularly interesting to assess the effects of more gratification for efforts and hard work in the Finnish system and the outcomes if girls' own stereotypical behaviors could be changed.

Concerning educational data mining and learning analytics, the work presented in this thesis concentrated on a higher level than on a school- or classroom-level, which constitutes the majority of the existing work in these disciplines (as discussed in **PII**). Further work will continue on the level of higher education management on a national level. In particular, an elaboration of article **PX** is currently under construction.

Finally, one of the most important directions for future work (which is—as discussed along with concluding **RQ1**—partly already in action) is related to ethics when leveraging real educational data for automatic decision making.

YHTEENVETO (FINNISH SUMMARY)

Automaattinen tietämyksen muodostaminen harvoista ja laajoista koulutuksellisista aineistoista - tapaus Suomi

Suomalainen koulutusjärjestelmä on saanut paljon julkisuutta 2000-luvulla. Erityisesti erinomaiset tulokset PISA-tutkimuksen (Programme for International Student Assessment - PISA) kolmella ensimmäisellä osallistumiskerralla tekivät peruskoulujärjestelmästä kansainvälisesti kuuluisan. Peruskoulutuksen kansallisia ominaispiirteitä on sen jälkeen tutkittu paljon, pääosin kasvatustieteen piirissä. Tutkimus on pohjautunut avoimesti saatavilla olevien aineistojen analysointiin ja koulutusjärjestelmän arvioinnin ja kehittämisen tukemiseen käyttäen laadullisia ja määrällisiä tutkimusmenetelmiä. Tässä väitöskirjassa tarkastellaan suomalaista koulutusjärjestelmää koskevan uuden tietämyksen muodostamista laajoista ja harvoista eli paljon puuttuvia arvoja sisältävistä aineistoista—erityisesti PISA-aineistosta—koulutuksellisen tiedonlouhinnan ja oppimisanalytiikan menetelmiä käyttäen ja niitä edelleen kehittäen. Työllä onkin kahdentyyppisiä tavoitteita: edistää tietämyksen muodostamisen menetelmiä, algoritmeja ja tulosten automaattista tulkittavuutta, sekä soveltaa kehitettyjä menetelmiä koulutuksen ilmiöiden parempaa ymmärrystä varten. Väitöskirja perustuu kymmeneen kansainväliseen julkaisuun, joista ensimmäisessä esitetään yleinen viitekehys koulutuksellisten aineistojen monipuolisen analysoinnin tueksi. Seitsemässä seuraavassa julkaisussa tarkastellaan ja kehitetään edelleen kvantitatiivisen tietämyksen muodostamisen menetelmiä, joissa huomioidaan PISA-aineistojen erityispiirteet. Kahdessa viimeisessä julkaisussa havainnollistetaan, kuinka suomalaisen koulutusjärjestelmän hallintaan liittyvää päätöksentekoa voidaan automatisoida ja parantaa aikaisemmin kehitettyjä menetelmiä ja viitekehyksiä hyödyntämällä. Kokonaisuudessaan työn tulokset tarjoavat uutta tietoa ja näkemyksiä suomalaisesta koulutus- ja korkeakoulujärjestelmästä.

REFERENCES

Abramo, G., D'Angelo, C. A., Soldatenkova, A., 2016. The dispersion of the citation distribution of top scientists' publications. Scientometrics 109 (3), 1711–1724.
URL http://dx.doi.org/10.1007/s11192-016-2143-7

Ackerman, M., Ben-David, S., Branzei, S., Loker, D., 2012. Weighted Clustering. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. pp. 858–863.

Adams, R., 2011. Comments on Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment. Retrieved from http://www.oecd.org/pisa/47681954.pdf.

Adams, R. J., Lietz, P., Berezner, A., 2013. On the use of rotated context questionnaires in conjunction with multilevel item response models. Large-scale Assessments in Education 1 (1), 1.

Aggarwal, C. C., Reddy, C. K., 2013. Data Clustering: Algorithms and Applications. CRC Press.

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD Record. Vol. 22. ACM, pp. 207–216.
URL http://www.almaden.ibm.com/cs/quest/papers/sigmod93.pdf

Ahlgren, P., Colliander, C., Persson, O., 2012. Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. Scientometrics 92 (3), 767–780.
URL http://www.akademiai.com/doi/abs/10.1007/s11192-012-0632-x

Ahlgren, P., Waltman, L., 2014. The correlation between citation-based and expert-based assessments of publication channels: SNIP and SJR vs. Norwegian quality assessments. Journal of Informetrics 8 (4), 985 – 996.
URL http://www.sciencedirect.com/science/article/pii/S1751157714000911

Allison, P. D., 2002. Missing data. Vol. 136 of Quantitative Applications in the Social Sciences. SAGE Publishing.

Alpaydin, E., 2010. Introduction to Machine Learning, 2nd Edition. The MIT Press, Cambridge, MA, USA.
URL https://mitpress.mit.edu/books/introduction-machine-learning

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recognition 46 (1), 243–256.

Arthur, D., Vassilvitskii, S., 2007. k-means++: The Advantages of Careful Seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp. 1027–1035.

Atzori, L., Iera, A., Morabito, G., 2010. The Internet of Things: A survey. Computer networks 54 (15), 2787–2805.

Auranen, O., Nieminen, M., 2010. University research funding and publication performance - An international comparison. Research Policy 39 (6), 822–834.
URL http://www.sciencedirect.com/science/article/pii/S0048733310000764

Äyrämö, S., 2006. Knowledge Mining Using Robust Clustering. Vol. 63 of Jyväskylä Studies in Computing. University of Jyväskylä.

Baker, R., 2015. Big data and education. Coursera.
URL https://fe68f43d-a-44087189-s-sites.googlegroups.com/a/tc.columbia.edu/bakeredmlab/moot/W001V001.pdf?attachauth=ANoY7cq8dywenY0cUcS5SFGvDZNqMrHp3nEFAJCZgaIeuPcRPu_mxV6ZMi6-X24GSorkzFZR9ggtmnPVM0sN2jpWH8LMCKNGp-oCpvLiqrW71zOIRwLwSQGCqdAg2NAoX7X23aIkh0ARl3xIZVUQ8HoKe2EGyNpQ-oMgZu5iKQKTKoiIfuCkTtWR-L9oIkFekr8juQKlPmzgsfuCDe2XOv6ZnqRqRLLV-g%3D%3D&attredirects=0

Baker, R. S., 2010. Data Mining for Education. International Encyclopedia of Education 7, 112–118.

Baker, R. S., Inventado, P. S., 2014. Educational Data Mining and Learning Analytics. In: Learning analytics. Springer, pp. 61–75.

Baker, R. S., Yacef, K., 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining 1 (1), 3–17.

Batista, G. E., Monard, M. C., 2003. An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence 17 (5-6), 519–533.

Bergner, Y., Colvin, K., Pritchard, D. E., 2015a. Estimation of Ability from Homework Items when There Are Missing and/or Multiple Attempts. In: Proceedings of the 5th International Conference on Learning Analytics and Knowledge. ACM, New York, NY, USA, pp. 118–125.
URL http://doi.acm.org/10.1145/2723576.2723582

Bergner, Y., Kerr, D., Pritchard, D. E., 2015b. Methodological Challenges in the Analysis of MOOC Data for Exploring the Relationship between Discussion Forum Views and Learning Outcomes. In: Proceedings of the 8th International Conference on Educational Data Mining. Educational Data Mining Society, pp. 234–241.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer, New York, New York, USA.

Bouchet, F., Kinnebrew, J. S., Biswas, G., Azevedo, R., 2012. Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. In: EDM. pp. 65–72.

Bramer, M., 2007. Principles of Data Mining. Undergraduate Topics in Computer Science. Springer.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., 2003. Statistical Modeling: The Two Cultures. Quality control and applied statistics 48 (1), 81–82.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and regression trees. CRC press.

Bryman, A., 2003. Triangulation. The Sage encyclopedia of social science research methods. Thousand Oaks, CA: Sage.

Bryman, A., 2004a. Secondary analysis and official statistics. Social research methods 2, 200–217.

Bryman, A., 2004b. Triangulation. In: The SAGE Encyclopedia of Social Science Research Methods. Sage Publications, Inc., pp. 1143–1144.
URL http://dx.doi.org/10.4135/9781412950589

Byrne, J. R., O'Sullivan, K., Sullivan, K., Feb 2017. An IoT and Wearable Technology Hackathon for Promoting Careers in Computer Science. IEEE Transactions on Education 60 (1), 50–58.

Calders, T., Pechenizkiy, M., 2012. Introduction to the special section on educational data mining. ACM SIGKDD Explorations Newsletter 13 (2), 3–6.

Carlson, R., Genin, K., Rau, M., Scheines, R., 2013. Student Profiling from Tutoring System Log Data: When do Multiple Graphical Representations Matter? In: Proceedings of the 6th International Conference on Educational Data Mining. pp. 12–19.

Celebi, E. M., Kingravi, H. A., 2012. Deterministic Initialization of the K-Means Algorithm Using Hierarchical Clustering. International Journal of Pattern Recognition and Artificial Intelligence 26 (07), 1–26.

Celebi, E. M., Kingravi, H. A., Vela, P. A., 2012. A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Systems with Applications 40 (1), 200–210.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., Thüs, H., 2012. A Reference Model for Learning Analytics. International Journal of Technology Enhanced Learning 4 (5–6), 318–331.

Cheema, J. R., 2014. A Review of Missing Data Handling Methods in Education Research. Review of Educational Research 84 (4), 487–508.

Chen, L., Chen, L., Jiang, Q., Wang, B., Shi, L., 2009. An Initialization Method for Clustering High-Dimensional Data. In: 1st International Workshop on Database Technology and Applications. IEEE, pp. 444–447.

Chin, J., Callaghan, V., July 2013. Educational Living Labs: A Novel Internet-of-Things Based Approach to Teaching and Research. In: 9th International Conference on Intelligent Environments. pp. 92–99.

Chu, W. W., 2014. Data Mining and Knowledge Discovery for Big Data. Vol. 1 of Studies in Big Data. Springer-Verlag Berlin Heidelberg.

Chua, A., 2011. Battle Hymn of the Tiger Mother. Bloomsbury Publishing.

Coughlan, S., 2016. Which country really has the cleverest students? http://www.bbc.com/news/business-37649892.

Creswell, J. W., 1994. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage publications.

Dash, M., Liu, H., 2000. Feature Selection for Clustering. In: Advances in Knowledge Discovery and Data Mining: 4th Pacific-Asia Conference, Proceedings. Springer, pp. 110–121.

Davies, D. L., Bouldin, D. W., 1979. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (2), 224–227.

Dawson, S., Gašević, D., Siemens, G., Joksimović, S., 2014. Current State and Future Trends: A Citation Network Analysis of the Learning Analytics Field. In: Proceedings of the 4th International Conference on Learning Analytics and Knowledge. ACM, pp. 231–240.

Denner, J., 2007. The Girls Creating Games Program: An Innovative Approach to Integrating Technology into Middle School. Meridian: A Middle School Computer Technologies Journal 1 (10).

Denzin, N., 1970. Strategies of Multiple Triangulation. The Research Act: A Theoretical Introduction to Sociological Methods, 297–313.

DiMaggio, P., 1997. Culture and cognition. Annual Review of Sociology, 263–287.

Domingos, P., Pazzani, M., 1997. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine learning 29 (2-3), 103–130.

Drezner, Z., Hamacher, H. W., 2001. Facility Location. Applications and Theory. Springer Science & Business Media.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7, 1–26.

Everitt, B., Skrondal, A., 2002. The Cambridge Dictionary of Statistics (Fourth Edition). Cambridge University Press.

Fairclough, R., Thelwall, M., 2015. More precise methods for national research citation impact comparisons. Journal of Informetrics 9 (4), 895 – 906.
URL http://www.sciencedirect.com/science/article/pii/S1751157715300894

Fayyad, U., Piatesky-Shapiro, S., Smyth, P., Nov. 1996a. Extracting Useful Knowledge from Volumes of Data. Communications of the ACM 39 (11), 27–34.
URL http://doi.acm.org/10.1145/240455.240464

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996b. From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press, pp. 1–30.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996c. From Data Mining to Knowledge Discovery in Databases. AI magazine 17 (3), 37.

Ferguson, R., 2012. Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4 (5–6), 304–317.

Ferguson, R., Cooper, A., Drachsler, H., Kismihók, G., Boyer, A., Tammets, K., Monés, A. M., 2015. Learning analytics: European perspectives. In: Proceedings of the 5th International Conference on Learning Analytics and Knowledge. ACM, pp. 69–72.

Ferguson, R., Shum, S. B., 2012. Social Learning Analytics: Five Approaches. In: Proceedings of the Second International Conference on Learning Analytics and Knowledge. ACM, pp. 23–33.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (2), 179–188.

Fix, E., Hodges, Jr., J. L., 1951. Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties. Tech. rep., DTIC Document.

Gaber, M., Zaslavsky, A., Krishnaswamy, S., 2005. Mining Data Streams: A Review. ACM Sigmod Record 34 (2), 18–26.

Gaber, S., Cankar, G., Umek, L. M., Tašner, V., 2012. The danger of inadequate conceptualisation in PISA for education policy. Compare: A Journal of Comparative and International Education 42 (4), 647–663.

Gifi, A., 1991. Nonlinear Multivariate Analysis. Wiley.

Gondek, D., Lally, A., Kalyanpur, A., Murdock, J. W., Duboué, P. A., Zhang, L., Pan, Y., Qiu, Z., Welty, C., 2012. A framework for merging and ranking of answers in DeepQA. IBM Journal of Research and Development 56 (3.4), 14–1.

Gray, G., McGuinness, C., Owende, P., Carthy, A., 2014. A Review of Psychometric Data Analysis and Applications in Modelling of Academic Achievement in Tertiary Education. Journal of Learning Analytics 1 (1), 75–106.

Gupta, D., Sharma, A., Unny, N., Manjunath, G., 2014. Graphical Analysis and Visualization of Big Data in Business Domains. In: Big Data Analytics, Lecture Notes in Computer Science (8883). Springer-Verlag, pp. 53–56.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. Adaptive Computation and Machine Learning. MIT Press.

Harpstead, E., MacLellan, C. J., Koedinger, K. R., Aleven, V., Dow, S. P., Myers, B. A., 2013. Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. In: Proceedings of the 6th International Conference on Educational Data Mining. pp. 51–58.

Haustein, S., Larivière, V., 2015. The Use of Bibliometrics for Assessing Research: Possibilities, Limitations and Adverse Effects. In: Incentives and Performance. Springer, pp. 121–139.
URL http://link.springer.com/chapter/10.1007%2F978-3-319-09785-5_8

Hawkins, W., Heffernan, N., Wang, Y., Baker, R. S., 2013. Extending the Assistance Model: Analyzing the Use of Assistance over Time. In: Proceedings of the 6th International Conference on Educational Data Mining. pp. 59–66.

Heller Sahlgren, G., 2015. Real Finnish Lessons. The True Story of an Education Superpower. London: Centre for Policy Studies.

Hershkovitz, A., Knight, S., Dawson, S., Jovanović, J., Gašević, D., 2016. About "Learning" and "Analytics". Journal of Learning Analytics 3 (2), 1–5.

Heyneman, S. P., Lee, B., 2013. The Impact of International Studies of Academic Achievement on Policy and Research. In: Handbook of International Large Scale Assessment: Background, Technical Issues, and Methods of Data Analysis. CRC Press, Taylor and Francis Group, LLC, pp. 37–74.

Hicks, D., 2012. Performance-based university research funding systems. Research Policy 41 (2), 251–261.
URL http://www.sciencedirect.com/science/article/pii/S0048733311001752

Hildebrandt, M., 2017. Learning as a Machine: Crossovers between Humans and Machines. Journal of Learning Analytics 4 (1), 6–23.

Hitlin, S., Piliavin, J., 2004. Values: Reviving a Dormant Concept. Annual Review of Sociology, 359–393.

Hofstede, G., 2011. Dimensionalizing cultures: The Hofstede model in context. Online readings in psychology and culture 2 (1), 8.

Hopmann, S., Brinek, G., Retzl, M., 2007. PISA According to PISA: Does PISA Keep what it Promises? Schulpädagogik und pädagogische Psychologie. Lit.
URL https://books.google.fi/books?id=f2m8PAAACAAJ

Huber, P. J., 2009. Robust Statistics (Second Edition). John Wiley & Sons Inc., New York, Wiley Series in Probability and Mathematical Statistics.

International Educational Data Mining Society, 2016.
URL http://www.educationaldatamining.org/

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31 (8), 651–666.

Jain, A. K., Dubes, R. C., 1988. Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data Clustering: A Review. ACM computing surveys (CSUR) 31 (3), 264–323.

Jauhiainen, S., Kärkkäinen, T., 2017. A Simple Cluster Validation Index with Maximal Coverage. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2017. pp. 293–298.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V., de Kereki, I. F., 2016. Translating Network Position into Performance: Importance of Centrality in Different Network Configurations. In: Proceedings of the 6th International Conference on Learning Analytics and Knowledge. ACM, New York, NY, USA, pp. 314–323.
URL http://doi.acm.org/10.1145/2883851.2883928

Jolliffe, I., 2002. Principal Component Analysis (2nd edition). Springer Series in Statistics. Springer New York.

Kärkkäinen, T., Äyrämö, S., 2004. Robust clustering methods for incomplete and erroneous data. WIT Transactions on Information and Communication Technologies 33.

Kärkkäinen, T., Heikkola, E., 2004. Robust formulations for training multilayer perceptrons. Neural Computation 16 (4), 837–862.

Kerr, D., Chung, G., 2012. Identifying key features of student performance in educational video games and simulations through cluster analysis. Journal of Educational Data Mining 4 (1), 144–182.

Kim, M., Ramakrishna, R., 2005. New indices for cluster validity assessment. Pattern Recognition Letters 26 (15), 2353–2363.

Kinnunen, P., Marttila-Kontio, M., Pesonen, E., 2013. Getting to know computer science freshmen. In: Proceedings of the 13th Koli Calling International Conference on Computing Education Research. Koli Calling '13. ACM, New York, NY, USA, pp. 59–66.
URL http://doi.acm.org/10.1145/2526968.2526975

Kivirauma, J., Ruoho, K., 2007. Excellence through special education? Lessons from the Finnish school reform. International Review of Education 53 (3), 283–302.

Kjærnsli, M., Lie, S., 2004. PISA and scientific literacy: similarities and differences between the nordic countries. Scandinavian Journal of Educational Research 48 (3), 271–286.
URL http://dx.doi.org/10.1080/00313830410001695736

Knodel, P., Windzio, M., Martens, K., 2014. Introduction: Outcomes and Actors–Reactions on Internationalization in Education Policy. In: Internationalization of Education Policy. Springer, pp. 1–34.

Kontkanen, P., Lahtinen, J., Myllymäki, P., Silander, T., Tirri, H., 2000. Supervised Model-based Visualization of High-dimensional Data. Intelligent Data Analysis 4 (3, 4), 213–227.

Koskela, A., 2016. Exploring the differences of Finnish students in PISA 2003 and 2012 using educational data mining. Jyväskylä Studies in Computing. University of Jyväskylä.

Kreiner, S., Christensen, K. B., 2014. Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. Psychometrika 79 (2), 210–231.

Kruskal, W., Wallis, W., 1952. Use of Ranks in One-Criterion Variance Analysis. Journal of the American statistical Association 47 (260), 583–621.

Kulczycki, E., Rozkosz, E. A., 2017. Does an expert-based evaluation allow us to go beyond the Impact Factor? Experiences from building a ranking of national journals in Poland. Scientometrics, 1–26.

Kyllonen, P. C., Bertling, J. P., 2014. Innovative Questionnaire Assessment Methods to Increase Cross-Country Comparability. In: Handbook of International Large Scale Assessment: Background, Technical Issues, and Methods of Data Analysis. CRC Press, Taylor and Francis Group, LLC, pp. 277–285.

Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Tech. rep., META Group.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lietz, P., Cresswell, J. C., Rust, K. F., Adams, R. D., 2017. Implementation of Large-Scale Education Assessments. Wiley.

Linnakylä, P., Välijärvi, J., I., A., 2011. Finnish Basic Education - When Equity and Excellence Meet. In: Equity and Excellence in Education: Towards maximal learning opportunities for all students. Routledge, New York, pp. 190–214.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data (2nd Edition). Wiley New York.

Liu, A., Xie, Y., 2014. Culture and Asian-White Achievement Difference. Population Studies Center Research Report 14-827. University of Michigan.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., 2010. Understanding of Internal Clustering Validation Measures. In: Proceedings of the 10th International IEEE Conference on Data Mining (ICDM). IEEE, pp. 911–916.

Looi, C.-K., Wong, L.-H., So, H.-J., Seow, P., Toh, Y., Chen, W., Zhang, B., Norris, C., Soloway, E., 2009. Anatomy of a mobilized lesson: Learning my way. Computers & Education 53 (4), 1120 – 1132, Learning with ICT: New perspectives on help seeking and information searching.
URL http://www.sciencedirect.com/science/article/pii/S036013150900133X

Marsman, M., 2014. Plausible Values in Statistical Inference. Universiteit Twente.

McCulloch, W. S., Pitts, W., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. The Bulletin of Mathematical Biophysics 5 (4), 115–133.

Meinck, S., 2015. Computing Sampling Weights in Large-Scale Assessments in Education. Survey Methods: Insights from the Field (SMIF).

Merceron, A., Blikstein, P., Siemens, G., 2016. Learning Analytics: From Big Data to Meaningful Data. Journal of Learning Analytics 2 (3), 4–8.

Merceron, A., Yacef, K., 2005. Clustering Students to Help Evaluate Learning. In: Technology Enhanced Learning. Springer US, Boston, MA, pp. 31–42.

Merz, C. J., Murphy, P. M., 1998. UCI Repository of Machine Learning Databases. URL {https://archive.ics.uci.edu/ml/datasets.html}

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.-R., 1999. Fisher Discriminant Analysis with Kernels. In: Proceedings of the IEEE Neural Networks for Signal Processing Workshop. IEEE, pp. 41–48.

Mislevy, R. J., 1991. Randomization-based inference about latent variables from complex samples. Psychometrika 56 (2), 177–196.

MOEC, 2012. Finnish Education in a Nutshell. Education in Finland. Ministry of Education and Culture (MOEC).
URL {http://www.oph.fi/download/146428_Finnish_Education_in_a_Nutshell.pdf}

Morgan, H., 2014. Review of Research: The Education System in Finland: A Success Story Other Countries Can Emulate. Childhood Education 90 (6), 453–457.

Musik, A., 2016. Philologenverband bezeichnet Pisa-Studie als Geldverschwendung. http://www.deutschlandfunk.de/bildungsforschung-in-der-kritik-philologenverband.680.de.html?dram:article_id=347675.

Niemi, H., Toom, A., Kallioniemi, A., 2012. Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools. Sense Publishers.

Niemi, H., Toom, A., Kallioniemi, A., 2016. Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools (Second Revised Edition). SensePublishers, Rotterdam.
URL http://dx.doi.org/10.1007/978-94-6300-776-4_1

Nokelainen, P., Silander, T., 2014. Using New Models to Analyze Complex Regularities of the World: Commentary on Musso et al. (2013). Frontline Learning Research 2 (1), 78–82.

OECD, 2009. PISA Data Analysis Manual: SPSS and SAS, Second Edition. OECD Publishing.

OECD, 2011. Finland: Slow and Steady Reform for Consistently High Results. In: Successful Reformers in Education: Lessons from PISA for the United States. OECD, pp. 117–135.

OECD, 2012a. Education at a glance: Highlights. OECD Publishing. Paris: France.
URL {http://dx.doi.org/10.1787/eag_highlights-2012-en}

OECD, 2012b. PISA 2009 Technical Report. OECD Publishing.

OECD, 2013a. PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II). PISA, OECD Publishing.
URL http://dx.doi.org/10.1787/9789264201132-en

OECD, 2013b. PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III). PISA, OECD Publishing.
URL http://dx.doi.org/10.1787/9789264201170-en

OECD, 2014a. PISA 2012 Results: What Students Know and Can Do (Volume I) Student Performance in Mathematics, Reading and Science: Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014). OECD Publishing.
URL http://dx.doi.org/10.1787/9789264201118-en

OECD, 2014b. PISA 2012 Technical Report.
URL https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD, 2015. Students, Computers and Learning: Making the Connection. OECD Publishing. Paris: France.
URL {http://dx.doi.org/10.1787/9789264239555-en}

OECD, 2017a. Mathematics performance (PISA) (indicator). Accessed on 20 February 2017.
URL {https://data.oecd.org/pisa/mathematics-performance-pisa.htm}

OECD, 2017b. Public spending on education (indicator). Accessed on 03 February 2017.
URL {https://data.oecd.org/eduresource/public-spending-on-education.htm}

Olsen, R. V., 2005a. Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science. Ph.D. thesis, University of Oslo.

Olsen, R. V., 2005b. An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension? Nordic Studies in Science Education 1 (1), 81–94.

Orlowski, P., 2017. Saskatchewan Teachers and a Study Abroad Experience in Finland: "I Love How the Finns Respect Their Teachers!". Journal of Educational Administration and Foundations 25 (3).

Pardo, A., Teasley, S., 2014. Learning Analytics Research, Theory and Practice: Widening the Discipline. Journal of Learning Analytics 1 (3), 4–6.

Pelánek, R., Rihák, J., Papousek, J., 2016. Impact of Data Collection on Interpretation and Evaluation of Student Models. In: Proceedings of the 6th International Conference on Learning Analytics and Knowledge. ACM, pp. 40–47.

Peña-Ayala, A., 2017. Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning. Springer International Publishing, Cham.

Piety, P. J., Hickey, D. T., Bishop, M., 2014. Educational Data Sciences - Framing Emergent Practices for Analytics of Learning, Organizations, and Systems. In: Proceedings of the 4th International Conference on Learning Analytics and Knowledge. ACM, pp. 193–202.

Pyle, D., 1999. Data preparation for data mining. Vol. 1. Morgan Kaufmann.

Quinlan, J. R., 2014. C4.5: Programs for Machine Learning. Elsevier.

Räty, H., Snellman, L., Mäntysaari-Hetekorpi, H., Vornanen, A., 1995. The parents satisfaction: Paternity elementary school activities and school reforms in the attitude. Education: Finnish Educational Research magazine 26 (1995): 3.

Ray, S., Turi, R. H., 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. In: Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. pp. 137–143.

Reich, J., Tingley, D. H., Leder-Luis, J., Roberts, M. E., Stewart, B., 2014. Computer-assisted reading and discovery for student generated text in massive open online courses. Journal of Learning Analytics 2 (1), 156–184.

Reinikainen, P., 2012. Amazing PISA results in Finnish comprehensive schools. In: Miracle of Education. Springer, pp. 3–18.

Rogers, T., 2015. Critical Realism and Learning Analytics Research: Epistemological Implications of an Ontological Foundation. In: Proceedings of the 5th International Conference on Learning Analytics and Knowledge. ACM, New York, NY, USA, pp. 223–230.
URL http://doi.acm.org/10.1145/2723576.2723631

Romero, C., Ventura, S., 2010. Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 40 (6), 601–618.

Rousseeuw, P. J., Leroy, A. M., 1987. Robust regression and outlier detection. John Wiley & Sons Inc., New York.

Rubin, D. B., 1976. Inference and missing data. Biometrika 63 (3), 581–592.

Rubin, D. B., 1987. Multiple Imputation for Nonresponse in Surveys. Tech. rep., JOHN WILEY & SONS.

Rutkowski, L., 2011. The Impact of Missing Background Data on Subpopulation Estimation. Journal of Educational Measurement 48 (3), 293–312.

Rutkowski, L., 2014. Sensitivity of Achievement Estimation to Conditioning Model Misclassification. Applied Measurement in Education 27 (2), 115–132.

Rutkowski, L., Gonzalez, E., Joncas, M., von Davier, M., 2010. International Large-Scale Assessment Data Issues in Secondary Analysis and Reporting. Educational Researcher 39 (2), 142–151.

Rutkowski, L., Rutkowski, D., 2010. Getting It "Better": The Importance of Improving Background Questionnaires in International Large-Scale Assessment. Journal of Curriculum Studies 42 (3), 411–430.

Rutkowski, L., Rutkowski, D., Zhou, Y., 2016. Item Calibration Samples and the Stability of Achievement Estimates and System Rankings: Another Look at the PISA Model. International Journal of Testing 16 (1), 1–20.

Saarela, M., Kärkkäinen, T., 2015. Analysing Student Performance using Sparse Data of Core Bachelor Courses. Journal of Educational Data Mining 7 (1), 3–32.

Saarela, M., Kärkkäinen, T., 2017. Knowledge Discovery from the Programme for International Student Assessment. In: Peña-Ayala, A. (Ed.), Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning. Springer International Publishing, Cham, pp. 229–267.

Saarela, M., Kärkkäinen, T., Lahtonen, T., Rossi, T., 2016. Expert-based versus citation-based ranking of scholarly and scientific publication channels. Journal of Informetrics 10 (3), 693 – 718.
URL http://www.sciencedirect.com/science/article/pii/S1751157715302194

Sahlberg, P., 2011. Finnish lessons. What can the world learn from educational change in Finland? Teachers College Press.

Sahlberg, P., 2015. Finnish lessons 2.0: What can the world learn from educational change in Finland? Teachers College Press.

Schafer, J. L., Graham, J. W., 2002. Missing Data: Our View of the State of the Art. Psychological methods 7 (2), 147.

Schatz, M., et al., 2016. Education as Finland's Hottest Export?: A Multi-Faceted Case Study on Finnish National Education Export Policies. Research Reports of the Department of Teacher Education.

Schleicher, A., 2007. Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? Journal of Educational Change 8 (4), 349–357.

Schwab, K., Sala-i Martín, X., Brende, B., 2013. The Global Competitiveness Report 2013–2014. World Economic Forum.

Siemens, G., 2013. Learning Analytics: The Emergence of a Discipline. American Behavioral Scientist 57, 1380–1400.

Siemens, G., 2014. The Journal of Learning Analytics: Supporting and Promoting Learning Analytics Research. Journal of Learning Analytics 1 (1), 3–5.

Siemens, G., Baker, R. S., 2012. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. ACM, pp. 252–254.

Simola, H., 2005. The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. Comparative education 41 (4), 455–470.

Simola, H., 2014. The Finnish Education Mystery: Historical and sociological essays on schooling in Finland. Routledge.

Singleton, Jr., R. A., Straits, B. C., Straits, M. M., 1993. Approaches to Social Research. Oxford University Press.

Spiegelhalter, D., 2013. PISA statistical methods - more detailed comments. https://understandinguncertainty.org/pisa-statistical-methods-more-detailed-comments.

Sprent, P., Smeeton, N. C., 2007. Applied Nonparametric Statistical Methods (Fourth Edition). CRC Press.

Springer, A., Johnson, M., Eagle, M., Barnes, T., 2013. Using sequential pattern mining to increase graph comprehension in intelligent tutoring system student data. In: Proceeding of the 44th ACM technical symposium on Computer science education. ACM, pp. 732–732.

Tan, C., 2017. Chinese responses to Shanghai's performance in PISA. Comparative Education 53 (2), 1–15.
URL http://dx.doi.org/10.1080/03050068.2017.1299845

Tan, P.-N., Steinbach, M., Kumar, V., 2007. Introduction to Data Mining. Pearson Education.

The Daily Telegraph, 2017. University degree subjects with the highest dropout rates. http://www.telegraph.co.uk/education/educationpicturegalleries/11002595/University-degree-subjects-with-the-highest-dropout-rates.html?frame=2824912, accessed: 2017-03-01.

Thorndike, R. L., 1953. Who belongs in the family? Psychometrika 18 (4), 267–276.

Tirri, K., Kuusisto, E., 2013. How Finland Serves Gifted and Talented Pupils. Journal for the Education of the Gifted 36 (1), 84–96.

Tirri, K., Nokelainen, P., 2012a. Ethical Thinking Skills of Mathematically Gifted Finnish Young Adults. in Talent Development & Excellence.

Tirri, K., Nokelainen, P., 2012b. Measuring Multiple Intelligences and Moral Sensitivities in Education. Vol. 5. Springer Science & Business Media.

Tölgyesi, C., Bátori, Z., Erdõs, L., 2014. Using statistical tests on relative ecological indicator values to compare vegetation units–Different approaches and weighting methods. Ecological Indicators 36, 441–446.

Tukey, J. W., 1977. Exploratory Data Analysis. Pearson.

Välijärvi, J., Kupari, P., Ahonen, A., Arffman, I., Harju-Luukkainen, H., Leino, K., Niemivirta, M., Nissinen, K., Salmela-Aro, K., Tarnanen, M., Tuominen-Soini, H., Vettenranta, J., Vuorinen, R., 2015. Millä eväillä osaaminen uuteen nousuun? PISA 2012 tutkimustuloksia.
URL {http://minedu.fi/export/sites/default/OPM/Julkaisut/2015/liitteet/okm6.pdf}

Välijärvi, J., Kupari, P., Linnakylä, P., Reinikainen, P., Sulkunen, S., Törnroos, J., Arffman, I., 2007. The Finnish success in PISA - and some reasons behind it: PISA 2003. University of Jyväskylä, Institute for Educational Research.

Välijärvi, J., Linnakylä, P., Kupari, P., Reinikainen, P., Arffman, I., 2002. The Finnish success in PISA - and some reasons behind it: PISA 2000. University of Jyväskylä, Institute for Educational Research.

Välijärvi, J., Sulkunen, S., 2016. Finnish School in International Comparison. In: Niemi, H., Toom, A., Kallioniemi, A. (Eds.), Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools (Second Revised Edition). SensePublishers, Rotterdam, pp. 3–21.
URL http://dx.doi.org/10.1007/978-94-6300-776-4_1

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., Karakos, A., 2012. A Clustering Methodology of Web Log Data for Learning Management Systems. Educational Technology & Society 15 (2), 154–167.

van Leeuwen, A., Janssen, J., Erkens, G., Brekelmans, M., 2015. Teacher regulation of cognitive activities during student collaboration: Effects of learning analytics. Computers & Education 90, 80 – 94.
URL http://www.sciencedirect.com/science/article/pii/S0360131515300439

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., Santos, J. L., 2013. Learning Analytics Dashboard Applications. American Behavioral Scientist 57 (10), 1500–1509.

Verleysen, M., Francois, D., 2005. The Curse of Dimensionality in Data Mining and Time Series Prediction. In: International Work-Conference on Artificial Neural Networks. Springer, pp. 758–770.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., et al., 1999. Self-organizing map in matlab: the som toolbox. In: Proceedings of the Matlab DSP Conference. Vol. 99. pp. 16–17.

von Davier, M., 2014. Imputing Proficiency Data under Planned Missingness in Population Models. In: Rutkowski, L., von Davier, M., Rutkowski, D. (Eds.), Handbook of International Large-Scale Assessment. Background, Technical Issues, and Methods of Data Analysis. CRC Press, Taylor and Francis Group, LLC, pp. 175–202.

Von Davier, M., Gonzalez, E., Mislevy, R., 2009. What are plausible values and why are they useful. IERI monograph series 2, 9–36.

von Davier, M., Sinharay, S., 2013. Analytics in international large-scale assessments: Item response theory and population models. In: Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis. CRC Press, Taylor and Francis Group, LLC, pp. 155–174.

Wagemaker, H., 2014. International Large-Scale Assessments: From Research to Policy. In: Handbook of International Large Scale Assessment: Background, Technical Issues, and Methods of Data Analysis. CRC Press, Taylor and Francis Group, LLC, pp. 11–36.

Waldow, F., Takayama, K., Sung, Y.-K., 2014. Rethinking the pattern of external policy referencing: media discourses over the "Asian Tigers" PISA success in Australia, Germany and South Korea. Comparative Education 50 (3), 302–321.
URL http://dx.doi.org/10.1080/03050068.2013.860704

Wallden, L. J., 2016. Kansainvälisten koulutusarvioiden vertailu koulutuksellisen tiedonlouhinnan keinoin. Jyväskylä Studies in Computing. University of Jyväskylä.

Waltman, L., van Eck, N. J., Wouters, P., 2013. Counting publications and citations: Is more always better? Journal of Informetrics 7 (3), 635–641.

Wang, Y., Paquette, L., Baker, R., 2014. A longitudinal study on learner career advancement in MOOCs. Journal of Learning Analytics 1 (3), 203–206.

White, H. D., 1990. Author co-citation analysis: Overview and defense. Scholarly communication and bibliometrics 84, 106.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Hill, J., Kain, R., Johnson, B., et al., 2015. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. Tech. rep., Higher Education Funding Council for England.
URL http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html

Witten, I., Frank, E., Hall, M., 2011. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

World Economic Forum, 2016. Global Gender Gap Report.
URL {http://www3.weforum.org/docs/GGGR16/WEF_Global_Gender_Gap_Report_2016.pdf}

Wu, M., 2005. The role of plausible values in large-scale surveys. Studies in Educational Evaluation 31 (2), 114 – 128.
URL http://www.sciencedirect.com/science/article/pii/S0191491X05000209

Ye, C., Biswas, G., 2014. Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information. Journal of Learning Analytics 1 (3), 169–172.

Young, D., 1954. Iterative Methods for Solving Partial Difference Equations of Elliptic Type. Transactions of the American Mathematical Society 76 (1), 92–111.

Zaki, M. J., Meira, Jr., W., 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

ORIGINAL PAPERS

PI

ANALYSING STUDENT PERFORMANCE USING SPARSE DATA OF CORE BACHELOR COURSES

by

Mirka Saarela, Tommi Kärkkäinen 2015

Journal of Educational Data Mining, 7(1):3–32

Analysing Student Performance using Sparse Data of Core Bachelor Courses

Mirka Saarela
University of Jyväskylä
mirka.saarela@jyu.fi

Tommi Kärkkäinen
University of Jyväskylä
tommi.karkkainen@jyu.fi

Curricula for Computer Science (CS) degrees are characterized by the strong occupational orientation of the discipline. In the BSc degree structure, with clearly separate CS core studies, the learning skills for these and other required courses may vary a lot, which is shown in students' overall performance. To analyze this situation, we apply nonstandard educational data mining techniques on a preprocessed log file of the passed courses. The joint variation in the course grades is studied through correlation analysis while intrinsic groups of students are created and analyzed using a robust clustering technique. Since not all students attended all courses, there is a nonstructured sparsity pattern to cope with. Finally, multilayer perceptron neural network with cross-validation based generalization assurance is trained and analyzed using analytic mean sensitivity to explain the nonlinear regression model constructed. Local (within-methods) and global (between-methods) triangulation of different analysis methods is argued to improve the technical soundness of the presented approaches, giving more confidence to our final conclusion that general learning capabilities predict the students' success better than specific IT skills learned as part of the core studies.

Keywords: Sparse Educational Data, Triangulation, Curricula Refinement, Correlation Analysis, Robust Clustering, Multilayer Perceptron

1. Introduction

The development of a curriculum for Computer Science (CS) can be challenging in an academic environment, given the discipline's strong occupational orientation. Especially at multidisciplinary universities (i.e., with many subject areas), the CS curriculum differs from the curricula of many other disciplines, as the core courses reflect to a large extent the vocational side of the program. In the case of the Department of Mathematical Information Technology (DMIT) at the University of Jyväskylä in Finland (reflecting both Finnish and European degree structures), the core bachelor courses compose only about 50 out of the minimum 180 ECTS (i.e., credits measured using the European Credit Transfer and Accumulation System) for the 3-year BSc degree (see Table 2). The degree contains other major courses in addition to separate introductory topics (e.g., general science, language and communication skills, statistics) and minor subject studies (especially mathematics). Students should acquire knowledge of very specific technical (e.g., programming) skills; however, computing interacts with many different domains, and in order to prepare students as the workforce of the future, domain knowledge as well as soft skills and personal attributes are important (Sahami et al., 2013a). For more than 40 years, roughly every 10 years, the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE) have promoted the creation of international curricular guidelines for bachelor programs in computing (Sahami et al., 2013b). Thus far, however, there has been little discussion about the relation between specific CS courses and other courses, in terms of the overall study performance.

Some researchers (see, e.g., Kinnunen et al. 2013 and references therein) indicate that primarily difficulties in mastering programming lead to high dropout rates in CS, therefore, one should pay special attention to them. Furthermore, a popular belief is that mathematical talent is the key skill for CS students to be successful (Jerkins et al., 2013). Although these topics are important, they do not cover the whole degree. The CS core of the DMIT curriculum for undergraduate students at the University of Jyväskylä, one of the largest and most popular multidisciplinary universities in Finland, has been more or less the same in recent years. Since the curriculum is typically updated every three years, the aim of this research is to focus on a set of mandatory courses related to the data collection period August 2009 through July 2013.

In addition, DMIT undergraduate students require more time to finish their studies compared to students of other disciplines at the University of Jyväskylä (Halonen, 2012). This happens even if the student's view on the quality of teaching and the study atmosphere at DMIT Jyväskylä is very positive and, in fact, better than in the whole Faculty of Information Technology (of which the DMIT is a part) or in the other departments at the university (Halonen, 2012). Actually, only a very few students (on average 12.8%) of DMIT complete the national target of at least 55 ECTS per academic year (Harden and Tervo, 2012). These study efficiency shortcomings apply to the absolute and relative number of credits and are especially important compared with students of other departments at the University of Jyväskylä, who amass many more credits in an academic year (29% acquire at least 55 ECTS).

To assess the current curriculum, we apply the educational data mining (EDM) approach. EDM consists of developing or utilizing data mining methods that are especially feasible for discovering novel knowledge originating in educational settings (Baker and Yacef, 2009) and supporting decision-making in educational institutions (Calders and Pechenizkiy, 2012). Most of the current case studies in EDM (see Table 1) analyze the steadily growing amount of log data from different computer-based learning environments, such as *Learning Management Systems* (e.g., Valsamidis et al. 2012), *Intelligent Tutoring Systems* (e.g., Hawkins et al. 2013; Bouchet et al. 2012; Carlson et al. 2013; Springer et al. 2013), or even *Educational Games* (e.g., Kerr and Chung 2012; Harpstead et al. 2013). Mining those data supports the understanding of how students learn and interact in such systems.

In our study, however, we are interested in understanding the effects of core CS courses and providing novel information for refining repetitive curricula. More specifically, we want to understand the effect of the current profile of the core courses on students' study success. These courses are taught in an ordinary fashion, meaning that in order to successfully complete a course, the student has to attend lectures, complete related exercises, and pass a final exam or assignment at the end of the course. The data analyzed in this paper are the historical log file from the study database at DMIT about all courses students passed for the period August 2009[1] until the end of July 2013. Patterns in our data provide improved profiling of the core courses and an indication of which study skills support timely and successful graduation.

[1] This is because since 8/2009 only ECTS-credits and separated Bachelor and Master degrees can be done.

The remainder of this paper is structured as follows. In Section 2, the overall methodology is explained. Section 3 is devoted to the correlation analysis, while in Section 4 we discuss our clustering analysis with robust prototypes. In Section 5, prediction analysis is realized with the multilayer perceptron (MLP) neural network. Conclusions from the domain as well as from the methodological level are presented in Section 6.

2. The overall methodology: Advocating multiphase triangulation

Baker et al. (2010) classify EDM methods into five categories: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. In Table 1, we summarize a representative set of EDM studies according to a) their data and the environment, b) goal of the study, c) EDM category and methods, and d) the knowledge discovered. This work was selected from forums, such as the Journal of Educational Data Mining, related annual conferences, and Google Scholar during autumn 2013. According to the table, which is organized by the different tasks and publication dates, scholars usually apply methods belonging to one of the classes of Baker et al.'s taxonomy to address a particular EDM problem. Moreover, predictive studies may apply many classifiers to assess the stability and reliability of the results. We, however, aim at multiphase triangulation: Different phases of the overall treatment within-methods and between-methods are varied and assessed (using rankings) to increase the technical soundness of the procedures and the overall reliability of the concluded results.

Generally, triangulation means that the same research objective is investigated by different data, theories, analysis methods, or researchers and then combined to arrive at convergent findings (Denzin, 1970). Probably the most popular way to apply triangulation is to use qualitative and quantitative methods and merge their results (Jick, 1979). We employ *between-method* triangulation (e.g., Denzin 1970; Bryman 2003), using techniques from distinct classes of the EDM taxonomy, to study the success patterns of the students who take the core courses of the computer science program in our department. First, we apply correlation analysis (Section 3), a key technique in *relationship mining*. Second, we utilize a special *clustering* approach (see Section 4) to find groups of students with similar course success. Third, we apply *prediction* (see Section 5) with model sensitivity analysis. In all between-methods, we discuss different *within-methods* that tighten the soundness of the respective between-method result. Moreover, we support our decision making a) in clustering with the *distillation of data for human judgement* (see our explorative and visual analysis in Section 4.2.1) and b) in prediction with *discovery with models* (model sensitivity is used as a component to calculate the mean variable sensitivity of the prediction model; see Section 5.1. To combine and interpret our results from the individual EDM techniques, we introduce a ranking system to which all the between and within analysis methods contribute.

In practice, the whole knowledge discovery process in our study is conducted by following the five classical stages (select the target data from the application domain, preprocess, transform, mine the transformed data, and interpret the results) introduced by Fayyad et al. (1996). Data preprocessing and transformation were performed in Java, while the data mining / machine learning techniques were either used as is (correlation analysis in Matlab's Mathematics package) or completely self-implemented (clustering and prediction as a whole) on the Mathworks Matlab R2013b platform.

Table 1: Overview of related work.

Environment and Data	Goal	**Category**: Methods	Obtained Knowledge
(San Pedro et al., 2013), United States (New York):			
Interaction data of a web-based tutoring system for mathematics from 3747 middle school students in New England plus college enrollment information for the students	Predict whether a student will (5 years later) attend college	**Prediction**: Logistic Regression Classifier	Students who are successful in middle school mathematics as measured by the tutoring system are more likely to enroll 5 years later in college, while students who are bored, confused, or careless in the system have a lower probability of enrolling.
(Vihavainen et al., 2013), Finland:			
Helsinki University, snapshot data from Computer Science student programming course	Predict whether a student will fail the introductory mathematics course	**Prediction**: Non-parametric Bayesian network tool (B-Course)	Students who cram at deadlines in their programming course are at high risk of failing their introductory mathematics course.
(Bayer et al., 2012), Czech Republic:			
Masaryk University, data of Applied Informatics bachelor students, their studies, and their activities in the university's information system (e.g., communication with other students via email/discussion board)	Predict whether a bachelor student will drop out of the university	**Prediction**: J48 decision tree learner, IB1 lazy learner, PART rule learner, SMO support vector machines, NB	Students who communicate with students who have good grades can successfully graduate with a higher probability than students with similar performance but who do not communicate with successful students.
(Kotsiantis, 2012), Greece:			
Hellenic Open University, data from distance learning course on Informatics	Predict students' final marks	**Prediction** M5', BP, LR, LWR, SMOreg, M5rules	Two written assignments predict the students' final grade the best.

Continued on next page

Table 1 – continued from previous page

Environment and Data	Goal	**Category**: Methods	Obtained Knowledge
(Bhardwaj and Pal, 2011), India:			
Purvanchal University, Department of Computer Applications, student data	Predict students' performance	**Prediction**: Bayesian Classifier	Living location has high influence on students' final grade.
(Mendez et al., 2008), United States (Arizona):			
Arizona State University, Science and Engineering student data	Prediction of student's persistence	**Prediction**: Decision Tree, Regression, Random Forest	High school and freshmen GPAs influence persistence the most.
(Erdogan and Tymor, 2005), Turkey:			
Maltepe University, data from student database	Find relations between performance on the entrance exam and later success	**Clustering**: K-means	The results of a student's university entrance exam determine the student's major in many cases.
(Campagni et al., 2012), Italy:			
University of Florence, Department of Computer Science, data of how and when exams were taken	Determine whether students who take exams in the recommended order are more successful	**Clustering**: K-means	Students who follow the *ideal path* perform better in terms of graduation time and final grade.
(Chandra and Nandhini, 2010), Nigeria:			
University in Nigeria, Department of Computer Science, course result data	Identify students' failure patterns	**Relationship Mining**: Apriori Association Rule Mining	Relationship between failed courses which can be used in order to restructure the curriculum (e.g., 2 introductory courses should be passed before the *Mathematical Modeling* course).

2.1. Data and nonstructured sparsity pattern

The original data, the historical log files of the four years, $8/2009 - 7/2013$, of all courses completed by all DMIT students, are challenging: Students are in different stages of their programs, their mandatory courses depend on their starting semester, they come with varying backgrounds,

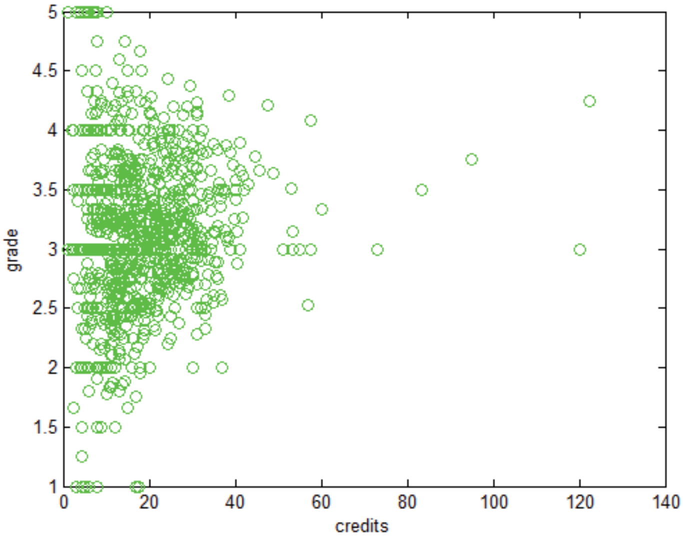


Figure 1: Relationship between the average credits per semester and grades.

have diverse interests, choose their optional courses accordingly, and, as a consequence, realize very different study profiles. This is a typical situation in multidisciplinary universities where students have the opportunity to choose from a large pool of courses. Altogether, our dataset consists of 13640 study records with 21 attributes, related to the passed course and the student's affiliation, and of 1040 students who attended a total of 1271 different courses, completing a total of 64905 credits. Only 64% of these credits the CS students obtained from courses in their own faculty.

When measuring the performance of individual students, in addition to quality, i.e., the grades, the quantity of studies, i.e., the number of all earned credits, is important. However, since our dataset consists of many students at different stages of their education, we cannot compare their individual sums of credits as is. Therefore, we assigned each passed course/record in our dataset to a semester, so that the *mean credits* (i.e., the average number of credits per student per semester) over the active semesters could be computed for all students. An active semester, in turn, is computed as the sum of all semesters between the first semester and the last semester that a student successfully completed a course. For example, a student who passed his or her first course in April 2010 and his or her last course in June 2013 has 7 active semesters. This may include semesters in which the student did not earn any credits. The *mean grade* is simply the sum of all grades divided by the number of courses a particular student has passed.

In general, quality and quantity of the studies of DMIT students do not correlate. The correlation coefficient between the average number of credits per student and the average grade is close to zero (0.0848). The per-student plot of the relationship between the number of credits per semester and the average grade is shown in Figure 1. Also the figure, which looks like a turned bell curve, which means that the grading of the courses resembles the normal distribution, shows visually that earned credits per student do not correlate with the average grade.

Table 2: Core bachelor courses.

course name	course code	course type	completion mode[2]	credits
Computer and Datanetworks as Tools	PCtools	introductory	assignment	2-4
Datanetworks	Datanet	introductory	exercises & final exam	3-5
Object Oriented Analysis and Design[3]	OOA&D	professional	exercises & final exam	3-6
Algorithms 1	Alg1	professional	final exam	4
Introduction to Software Engineering	IntroSE	professional	final exam	3
Operating Systems	OpSys	professional	final exam	4
Basics of Databases and Data Management	DB&DMgm	professional	final exam	4
Programming 1	Prog1	programming	assigment & final exam	6
Programming 2	Prog2	programming	assigment & final exam	8
Computer Structure and Architecture	CompArc	introductory	exercises & final exam	3
Programming of Graphical User Interfaces	GUIprog	programming	exercises & final exam	5
Research Methods in Computing	CompRes	methodological	essay	2
All core courses				47-54

Our goal is to better understand the students' success patterns, given the core courses, in relation to the rest of their studies. Therefore, we want to analyze the students who have completed a certain percentage of the courses of interest. The core courses, a specific set of 12 courses that, for that period of time we study, have been a mandatory part of the curriculum for all DMIT bachelor students, are listed and characterized in Table 2.

If we transform our data in such a way that the 12 core courses become the variables and the attribute value of each observation, corresponding to one student, is the grade of the core course or *missing* if the student did not attend or pass the course, the assembled matrix is very sparse. Only for 13 students are the rows full; the students have passed all the core courses. In Table 3, the high percentage of missing values and the sparsity of the matrix are summarized. The table shows how many students have completed exactly, and respectively at least, q of the 12 courses. Moreover, in each case the percentage of missing values of the cumulative data matrix is provided. The missing data values in the matrix are *missing at random* (Rubin, 1976; Rubin and Little, 2002). This means that the missing values are related to particular variables (some courses that are usually taken later in the program are completed by fewer students; see Figure 2) but not missing because of the values (grades) that could be observed if a particular course is passed.

To analyze such data, one cannot accept too many missing values. In this respect, the *breakdown point* related to statistical estimates (see, e.g., Hettmansperger and McKean 1998) on how much contamination (errors, missing values) in data can be tolerated is informative. An upper bound is easy to establish: If more than 50% of data is missing, then "missing" is the most typical value (mode) of the data. Furthermore, tests conducted with synthetic data show that,

[2] The difference between *assignment* and *exercises* in our system is important: While *assignment* denotes a mandatory work that the student has to fulfill in order to pass the course and affects the final grade the student will receive, *exercises* are smaller (usually weekly) optional tasks that correspond to the current lecture material.

[3] In spring 2012, the *Object Oriented Analysis and Design* (*OOA&D*) course was split into two separate courses, *Object Oriented Analysis* and *Object Oriented Design*. Therefore, in further analysis the following strategy was applied: If a student completed the original *Object Oriented Analysis and Design* course, the grade from this course was taken for the analysis. However, in case the student did not attend the original course, we used the mean grade of the *Object Oriented Analysis* and the *Object Oriented Design* course as the grade for *OOA&D* if the student had completed both newly created courses, or just the grade of the one course if the student had completed only one of these two courses.

Table 3: Number of students who have completed exactly q (n_q) or at least q ($\sum_{q=12}^{Q} n_q,\ Q = 12, \ldots, 0$) of the core courses during the analyzed period.

q	n_q	$\sum n_q$	missing values
12	13	13	0.0%
11	16	29	4.56%
10	22	51	9.81%
9	26	77	14.93%
8	49	100	19.17%
7	28	128	24.10%
6	35	163	29.65%
5	44	207	35.75%
4	46	253	41.37%
3	40	293	45.96%
2	82	375	54.13%
1	126	501	63.57%
0	539	1040	82.45%

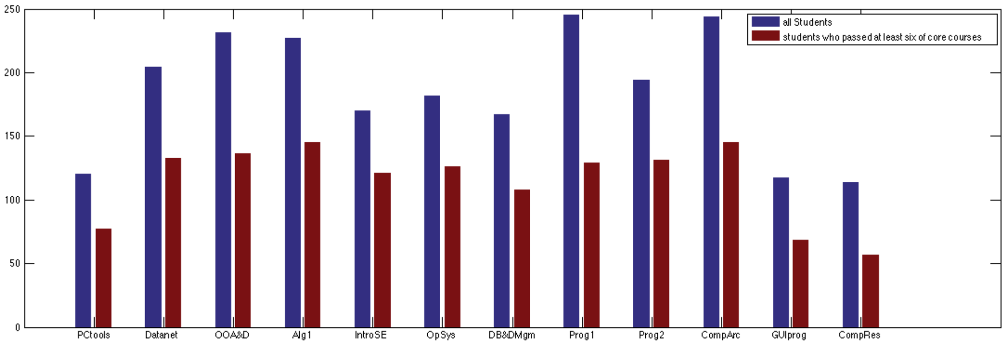


Figure 2: Number of students who passed coursewise.

for example, in clustering with robust methods, reliable results, i.e., almost zero error, can be obtained even if around 30% of the data is missing (Äyrämö 2006; see in particular Figure 22 at page 131). Therefore, our *data selection strategy* is to use that part of the whole, sparse data matrix, which contains the students who have completed at least half of the core courses. This dataset has about 30% missing values (see Table 3) for the multivariate techniques. In the correlation analysis (see Section 3), where the courses are analyzed individually, we similarly use the subsets of the students who have passed the particular course and at least five other courses additionally. In addition, different subsets of the sparse study matrix are utilized to realize some parts of cluster analysis and predictive analysis procedures.

A further challenge, particularly for predicting the study success (see Section 5), is that, for our primary target group, the number of credits related to the core courses is typically less than half of the total number of the earned credits. In Table 4, the percentages of credits originating from the core courses in relation to the total number of credits for the 163 students of interest (see Table 3) are shown. As can be seen in the table, for more than 70% of the students, the core courses account for fewer than half of their studies.

Table 4: Binning of students (nr = number) according to means of the number of core courses in relation to whole studies.

% core courses	nr	cumulative (%)
0-10%	16	16 (10%)
10-20%	24	40 (25%)
20-30%	24	64 (39%)
30-40%	29	93 (57%)
40-50%	25	118 (72%)
50-60%	17	135 (83%)
60-70%	15	150 (92%)
70-80%	7	157 (96%)
80-90%	4	161 (99%)
90-100%	2	163 (100%)

Summing up, for our analysis we have the entire base of completed courses (1040x21) that is processed and transformed to further subsets and the sparse 163x12 data matrix of the students who have completed at least half of the core courses and the grades they received in these courses.

3. Correlation analysis with Bonferroni correction

As our first EDM technique, we apply relationship mining using correlation analysis. In general, we know from Figure 1 that in terms of grades well-scoring students are not necessarily more likely to study actively. But how about the correlation for our target group, those students who have already completed at least half of the core courses? In the correlation analysis, we do not need special methods for the sparse data. However, the number of students who have passed an individual course differs considerably (see Figure 2) so that the correlation coefficients are computed for different student subsets. The mean number of credits and the mean grade are computed in the same way as explained in Section 2.1.

In Table 5, the correlation of each core course to (i) the mean grade of a student (denoted as *corr.grades*) and (ii) the mean number of credits per semester (denoted as *corr.credits*) is summarized. In each case, *r* identifies the calculated correlation, and *p* corresponds to the *p-value* for testing the hypothesis of no correlation, respectively. The number of stars indicates the strength of the evidence for no correlation. As usual, $\star$ symbolizes the *borderline to be significant* ($p <= 0.05$), $\star\star$ symbolizes *statistically significant* ($p <= 0.01$), and $\star\star\star$ symbolizes *highly statistically significant* ($p <= 0.005$). *rank* denotes the ordering of courses by means of the computed correlations.

From Table 5, we can conclude that, except the *Research Methods in Computing*, all courses have a moderate positive linear relationship to the students' general study success. The course-wise correlations to mean credits per semester are all positive as it should be (passing a course increases credits). In addition, all *corr.grades* illustrate that students who score high in those courses tend to score high in their other courses as well. In particular, this applies to four courses: *Algorithms 1*, *Computer Structure and Architecture*, *Datanetworks*, and *Programming 2*. The correlation between the grades for these four courses and the average grade of the student is in all cases highly statistically significant as the p-values for testing the hypothesis of no correlation are all smaller than 0.005. Similarly as with the classical p-test, we obtained with the conserva-

Table 5: Correlation of each core course to the students' general performance.

	corr.grades				corr.credits			
Course Code	*r*	*p*	*Bonferroni*	*rank*	*r*	*p*	*Bonferroni*	*rank*
PCtools	0.4164	⋆⋆⋆	⋆⋆⋆	11	0.1058	–	–	12
Datanet	0.6244	⋆⋆⋆	⋆⋆⋆	**3**	0.3887	⋆⋆⋆	⋆⋆⋆	**1**
OOA&D	0.5346	⋆⋆⋆	⋆⋆⋆	6	0.1327	–	–	11
Alg1	0.6593	⋆⋆⋆	⋆⋆⋆	**1**	0.3082	⋆⋆⋆	⋆⋆⋆	**4**
IntroSE	0.4197	⋆⋆⋆	⋆⋆⋆	10	0.1717	–	–	9
OpSys	0.5113	⋆⋆⋆	⋆⋆⋆	8	0.1905	⋆	–	8
DB&DMgm	0.5572	⋆⋆⋆	⋆⋆⋆	5	0.3312	⋆⋆⋆	⋆⋆	5
Prog1	0.4314	⋆⋆⋆	⋆⋆⋆	9	0.2438	⋆⋆	–	7
Prog2	0.5731	⋆⋆⋆	⋆⋆⋆	**4**	0.3549	⋆⋆⋆	⋆⋆⋆	**2**
CompArc	0.6511	⋆⋆⋆	⋆⋆⋆	**2**	0.3216	⋆⋆⋆	⋆⋆⋆	**3**
GUIprog	0.5343	⋆⋆⋆	⋆⋆⋆	7	0.3054	⋆	–	6
CompRes	0.2543	–	–	12	0.1609	–	–	10

tive *Bonferroni correction* (Rice, 1989) that the correlation of all core courses to the student's overall grade (except the *Research Methods in Computing*) are highly statistical relevant.

Another conclusion that can be made from Table 5 is that the same four courses that have the highest correlations to the general success of the student also have the highest correlation to the average number of credits. This means that if a student gets a high grade in these courses he or she will probably earn, on average, a high number of credits in the semester as well. Again, all of these findings are, according to the classical p-test as well as the Bonferroni correction, highly statistically significant. Although the ranking is different (e.g., while *Algorithm 1* correlates the most with the mean grade for the student, *Datanetworks* correlates the most with the mean number of credits per semester), we can conclude that those four courses correlate with the students' general performance the best.

To sum up, it can be inferred that a student who achieves a high grade in *Algorithms 1*, *Computer Structure and Architecture*, *Datanetworks*, or *Programming 2* is likely to be successful in the remaining part of his or her studies not only with the grade level but also in terms of speed of completing courses. Albeit overall semesterwise credits and average grade do not correlate at all (see Figure 1), a linear dependency between the grades a student received in the core courses and the general performance exists.

4. Cluster analysis using robust prototypes

Our second EDM method is clustering. Generally, clustering can be divided into *partitional* and *hierarchical* clustering (Jain, 2010; Steinbach et al., 2004). However, hierarchical clustering is appropriate only in very small datasets since most of the hierarchical algorithms have quadratic or higher computational complexity (Emre Celebi et al., 2012). Partitional clustering, however, is very efficient and scalable. It partitions the data, such that similar observations are assigned to the same subset of data (referred as a cluster), each observation is attributed to exactly one subset, and each subset contains at least one observation. Since we want to obtain a directly interpretable result, prototype-based partitional clustering is an appropriate approach here. If we can find a partition of data, where each cluster is represented by exactly one prototype, we can use this prototype to analyze the corresponding cluster. Prototype-based partitional clustering

Algorithm 1: Iterative relocation clustering algorithm

```
Input: Dataset and the number of clusters $K$.
Output: $K$ partitions of the given dataset.
Select $K$ points as the initial prototypes;
repeat
    1. Assign individual observation to the closest prototype;
    2. Recompute the prototypes with the assigned observations;
until The partition does not change;
```

can be realized using the iterative relocation algorithm skeleton presented in Algorithm 1 with different score functions (Han et al., 2001) according to which the two steps inside the loop of Algorithm 1 are optimized.

However, in order to realize a prototype-based partitive clustering algorithm, two main issues should be addressed. First, a well-known problem of all iterative relocation algorithms is their initialization. They minimize the given score function locally by iteratively relocating data points between clusters until an optimal partition is attained. Therefore, basic iterative algorithms, such as K-means, always converge to a local, and not necessarily to the global, optimum. Although much work has focused this problem, no efficient and universal method for identifying the initial partitions and the number of clusters exists. This problem is discussed more thoroughly in Section 4.2. The second problem is the sparse student data with around 30% missing values (see Section 2.1). In Section 4.1, a solution is presented for adjusting the score function of the basic algorithm skeleton in order to deal with the random sparsity pattern. A similar approach was also applied in Saarela and Kärkkäinen (2014) to other educational data.

4.1. Score function for K-spatialmedians

Our (available) data consist of course grades of fixed values $\{1, 2, 3, 4, 5\}$. Therefore, there is evidently a significant quantization error from uniform distribution in the probability distribution for a grade g_i:

$$g_i(x) = \begin{cases} 1, \text{if } g_i - \frac{1}{2} \leq x < g_i + \frac{1}{2}, \\ 0, \text{elsewhere.} \end{cases} \tag{1}$$

Thus, second-order statistics that rely on the normally distributed error are not suitable here, and we need to use the so-called nonparametric (i.e., robust) statistical techniques (Huber, 1981; Rousseeuw and Leroy, 1987; Hettmansperger and McKean, 1998). The simplest of robust location estimates are the median and the spatial median. The median, i.e., the middle value of the ordered univariate sample, is inherently one-dimensional, and thus with missing data uses only the available values of an individual variable. The spatial median, however, is truly a multidimensional location estimate and can take advantage of the available data pattern as a whole. This is illustrated and more thoroughly explained in Kärkkäinen and Heikkola (2004); especially, formulae (2.8) and (2.9) and Figures 1 and 2. As stated, e.g., in Croux et al. (2010), the spatial median is not affine but only orthogonally equivariant. However, because we have the fixed grade scale, this property of a statistical estimate is not necessary here. Moreover, for elliptical distributions, this behavior creates more scatter than location estimation (Croux et al., 2010). As a whole, the spatial median has many attractive statistical properties. In particular, its breakdown point is 0.5; it can handle up to 50% of contaminated data, which makes the spatial median very appealing for high-dimensional data with severe degradations and outliers. A missing value can be thought of as an infinite outlier because it can have any value (from the value range).

Äyrämö (2006) introduced a robust approach utilizing the spatial median to cluster very sparse and apparently noisy data: The *K-spatialmedians* clustering algorithm is based on the same algorithm skeleton as presented in Algorithm 1 but uses the projected spatial median as a score function:

$$\mathcal{J} = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \| \operatorname{diag}\{\boldsymbol{p}_i\}(\boldsymbol{x}_i - \boldsymbol{c}_j)\|_2, \tag{2}$$

Here, diag transforms a vector into a diagonal matrix. The latter sum in (2) is computed over the subset of data attached to cluster j and the projection vectors $\boldsymbol{p}_i, i = 1, \ldots, N$, capture the existing variable values:

$$(\boldsymbol{p}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases}$$

In Algorithm 1, the projected distance as defined in (2) is used in the first step, and recomputation of the prototypes, as the spatial median with the available data, is realized using the sequential overrelaxation (SOR) algorithm (Äyrämö, 2006) with the overrelaxation parameter $\omega = 1.5$. In what follows, we refer to Algorithm 1 with the score function (2) as *K-spatialmedians* clustering.

4.2. Initialization

It is a well-known problem that all iterative clustering algorithms are highly sensitive to the initial placement of the cluster prototypes, and thus, such algorithms do not guarantee unique clustering (Meilă and Heckerman, 1998; Emre Celebi et al., 2012; Bai et al., 2012; Jain, 2010). One might even argue that the results are not reliable if the initial prototypes are randomly chosen since the algorithms do not converge to a global optimum. Numerous methods have been introduced to address this problem. Random initialization is still often chosen as the general strategy (Xu and Wunsch, 2005). However, several researchers (e.g., Aldahdooh and Ashour 2013; Bai et al. 2011) report that having some other than random strategy for the initialization often improves final clustering results significantly.

An important issue when clustering data and finding an appropriate initialization method is the definition of (dis-)similarity of objects. Bai et al. (2011) and Bai et al. (2012) proposed initialization methods for categorical data. The attribute values of our dataset (grades from 1-5, or missing) are also categorical. However, the ordering of our attribute values has meaning (ordinal data). For example, a student who received grade 5 in all his or her courses is more dissimilar to a student who got mostly grade 2 than to a student who received on average grade 4. Therefore, an initialization method for data where only enough information is given to distinguish one object from another (nominal data) might not be suitable for our case.

Chen et al. (2009) proposed a novel approach to find good initial prototypes. Chen et al. argue that in the high-dimensional space data are inherently sparse. Therefore, the distance between each pair of observations becomes almost the same for a wide variety of data distributions. However, this approach seems more suitable for very high-dimensional data than for our 12-dimensional case. Emre Celebi et al. (2012) compared different initialization methods. They conclude that for small datasets (fewer than 10000 observations) Bradley and Fayyad's method

Algorithm 2: Constructive initialization approach for robust clustering

```
Input: Datasets $\mathbf{D_0}$ to $\mathbf{D_6}$.
Output: The set of prototypes for every value of $K$
for $K = size(D_0)$ to $2$ do
    $KBestPrototypes$ = globalBestSolution($D_0$,$K$);
    for $p = 1$ to $6$ do
        $KBestPrototypes$ = K-spatialmedians($D_p$,$K$,$KBestPrototypes$);
    end
end
```

leads to best results. In Bradley and Fayyad's method (1998), the original dataset is first split into smaller subsets that themselves are clustered. Then the temporary prototypes obtained from clustering the subsets are combined and clustered as many times as there are different subsets. Thus, each time one different set of temporary prototypes is tried as initialization and the best, i.e., that set of temporary prototypes which resulted in the smallest clustering error, is finally used as initialization for clustering the original dataset.

To sum up, the ideal approach for computing initial prototypes depends on the data, and is therefore context dependent. However, some general criteria apply: First, initial prototypes should be as far from each other as possible (Khan and Ahmad, 2013; Jain, 2010). Second, outliers or noisy observations are not good candidates as initial prototypes. Moreover, for relatively small datasets it seems to be a good idea to further divide the set into subsets and utilize the best prototypes of the smaller sets for further computations. Furthermore, as pointed out by Bai et al. (2012), it is advantageous if at least one initial prototype is close to a real solution. Bearing these issues in mind, we developed a new deterministic and context-sensitive approach to find good initial prototypes.

4.2.1. Initialization for sparse student data

Our intention is to interpret and characterize each cluster by its prototype. Therefore, we should prefer full prototypes, those that have no missing values. For this approach, we first note that the rows of Table 3 represent cascadic (see Kärkkäinen and Toivanen 2001) sets of data. Let us denote the datasets as D_p with $p = 0 \ldots 12,$ where $p = q - 12$. Thus, D_0 represents the very small but full dataset with the 13 students who have completed all 12 core courses and D_1 the 29 students who have completed at least 11 of them (containing D_0). Therefore, in general D_p consists of students who have passed exactly $12 - p$ of the core courses, and we always have $D_{p-1} \subset D_p$. This creates the basis for the proposed initialization approach, which is depicted as a whole in Algorithm 2.

Our initial, the complete dataset D_0 is so small that we can easily determine the globally best solution by minimizing the error of the spatial median by testing all possible initializations for the values of K[4] In Algorithm 2, *globalBestSolution* refers a function that tests all possible K combinations of the observations in the small complete dataset and returns the prototypes of the combination that resulted in the smallest clustering error. In that way, we obtain for every K for our small dataset the K global best prototypes. We then use the K best prototypes (denoted as $KBestPrototypes$ in the algorithm) on D_p as the initial prototypes for the next larger dataset

[4] Even if K is unknown, we can assume that K is at least 2 and smaller than the total number of observations.

Table 6: Comparison of context-sensitive and random initialization for robust clustering.

	context-sensitive		random	
K	error	missing values	error	missing values
13	373.54	15.38%	425.46	51.54%
12	374.04	8.33%	429.26	46.67%
11	372.31	0.00%	432.89	38.18%
10	376.57	0.00%	434.51	44.00%
9	391.42	0.00%	436.98	38.89%
8	396.06	0.00%	434.77	33.75%
7	409.70	0.00%	444.50	31.43%
6	425.93	0.00%	443.27	18.33%
5	437.32	0.00%	452.74	16.00%
4	454.27	0.00%	461.71	10.00%
3	471.79	0.00%	480.99	6.66%
2	506.84	0.00%	515.06	5.00%

D_{p+1}. Thus, throughout the constructive approach full prototypes and small clustering error are favored. The dataset D_6, the students who have completed at least half of the core courses, is our actual target data for clustering.

In Table 6, it is shown how the score function changes and the number of missing values with the proposed initialization strategy for different values of K for D_6. For comparison, the table also shows the average results of 10 test runs of the *K-spatialmedians* algorithm with random initialization. We obtain better results with our approach for the clustering error and, especially, with respect to the missing values. For example, already for $K = 3$, 6.66% of the prototypes' values are missing with random initialization and, thus, uninterpretable. Moreover, we also studied the stability of the results by checking whether the students in $D_{p-1},\ p \geq 1$, still belong to the same cluster when new students are added and the reclustering of D_p is performed in Algorithm 2. Confusion matrices between the two consecutive clustering levels were computed. It turned out that the confusion matrices are almost perfect, so that the formation of clusters is very stable and the clusters themselves are reliably structured. We conclude that the proposed context-sensitive initialization provides a clustering result with low error and high interpretability.

The best value for *K* is next determined using visual inspection. To avoid overfitting, our goal is to have a small number of clusters. However, the observations should not be too far away from the prototype to which they belong. From Figure 3, the plot of the second column in Table 6 (change in the score function when D_6 is clustered using the proposed strategy), we conclude that $K = 3, K = 5$, and $K = 8$ are potential values for the number of clusters. Namely, after precisely these points, the speed of the decrease (improvement) of the clustering error, the discrete derivative, slows down slightly (see, e.g., Zhong et al. 2008 for a similar approach). Of the potential values, we choose the first one that provides the smallest number of clusters for further analysis and, in such a way, generalizes data the most. The prototypes for $K = 3$ are visualized in Figure 4.

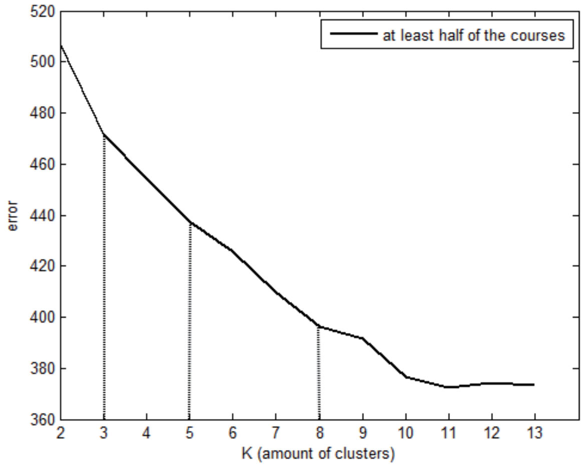


Figure 3: Decrease in errors for target data when more clusters are introduced successively.

4.3. Analyzing the clustering results

In the first two columns of Table 7, the ranking of the core courses based on their prototype separation is provided. Since the general profile of the three clusters is “medium” (*cluster 1*), “high” (*cluster 2*), and “low” (*cluster 3*), we compute, for each variable, two distances: $d_1 = |C_2 - C_1|$ and $d_2 = |C_3 - C_1|$. The two measures are computed (i) as the mean of $\{(d_1)_i, (d_2)_i\}$ (denoted as *measure 1*) and (ii) the minimum of $\{(d_1)_i, (d_2)_i\}$ (denoted as *measure 2*). As can be seen from the table, *measures 1* and *2* provide practically the same ranking. However, we think that of these two indicators, the second measure provides clearer variable separation. For example, with *measure 1* we could have a high distance value for a course even if only one prototype value C_i is very dissimilar from the other two. Moreover, in order to assess even further the explanative power of variables related to the clustering result with $K = 3$, we also applied the nonparametric Kruskal-Wallis test (Hollander et al., 2013) to compare the subsets of data in the three clusters. Since the actual clusterwise datasets contain missing values, we used one iteration of the *hot deck imputation* (Äyrämö, 2006; Batista and Monard, 2003) to complete them: We imputed the missing values using the cluster prototype values of the *K-spatialmedians* algorithm (see Section 4.1) with eight clusters (see Figure 3). As concluded in Section 4.2.1, eight was another good value for the number of clusters *K*. Because of this imputation and because of the form of the quantization error as explained in connection with formula (1), a nonparametric test should be used. According to the Kruskal-Wallis test, the difference between the different clusters is highly statistically significant for all courses. Again, the same four courses provide the highest differentiation between the clusters (see third column of Table 7) with only the *Operating Systems* very different from the distance-based separators. In the fourth column of Table 7, the sum of ranks from the second distance measure and the Kruskal-Wallis test are given, and the overall ranking of the courses based on the sum is provided. This rank-of-rankings approach is an example of within-method triangulation, where the final order of importance combines assessments of the prototypes and the clusterwise data subsets.

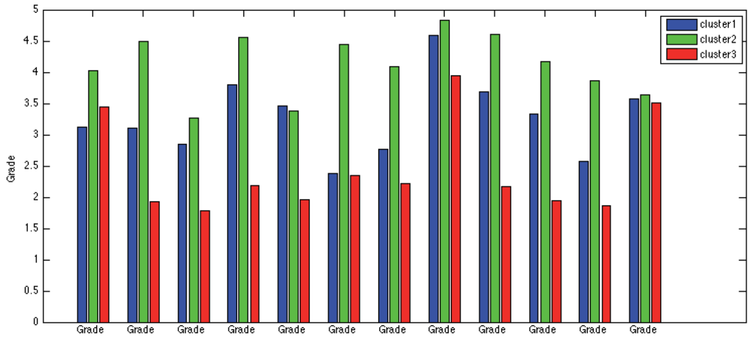


Figure 4: Prototypes of the three student clusters.

Table 7: Distances between the clusters.

	measure 1		measure 2		Kruskal-Wallis			
course code	distance	*rank*	distance	*rank*	χ^2	*p*	*rank*	sum (*rank*)
PCtools	0.1472	8	0.3220	8	33.10	$\star\star\star$	9	17 (9)
Datanet	0.8183	1	1.1754	1	84.69	$\star\star\star$	1	2 (**1**)
OOA&D	0.2249	7	0.4247	7	47.97	$\star\star\star$	7	14 (6)
Alg1	0.6090	3	0.7501	4	82.39	$\star\star\star$	2	6 (**2**)
IntroSE	0.0588	10	0.0781	10	34 .94	$\star\star\star$	8	18 (10)
OpSys	0.0413	11	0.0402	12	67.06	$\star\star\star$	4	16 (7)
DB&DMgm	0.3666	6	0.5490	6	53.96	$\star\star\star$	6	12 (5)
Prog1	0.0796	9	0.2417	9	32.21	$\star\star\star$	10	19 (11)
Prog2	0.7064	2	0.9309	2	54.35	$\star\star\star$	5	7 (**4**)
CompArc	0.5872	4	0.8439	3	78.49	$\star\star\star$	3	6 (**3**)
GUIprog	0.4602	5	0.7118	5	31.04	$\star\star\star$	11	16 (8)
CompRes	0.0021	12	0.0633	11	17.23	$\star\star\star$	12	23 (12)

The first observation that can be made by comparing the overall cluster ranking with the correlation analysis (see Table 5) is that the correlations are reflected in the different clusters. The four courses with the highest correlations clearly separate the three clusters. This can be seen as well from the visualization of the cluster prototypes (Figure 4). The students in the lowest-performing *cluster 3* also have the lowest performance in the *Datanetworks* and *Algorithms 1* course. The prototype of the best *cluster 2* is represented by a remarkable higher grade for those courses. The same applies for the other two courses with a high correlation to the average grade and the average number of credits of the students, *Computer Structure and Architecture* and *Programming 2*. A second interesting observation is that one of the smallest deviations in the grade is obtained for the *Research Methods in Computing*. This was also the only course that did not show a significant correlation to the students' overall grade (see Section 3). However, also the content of this course differs from the other core courses by being not directly related to IT knowledge. Moreover, in contrast to all other courses, this course is evaluated solely by an essay that the student has to write (see Table 2). Somewhat exceptional behavior for this course

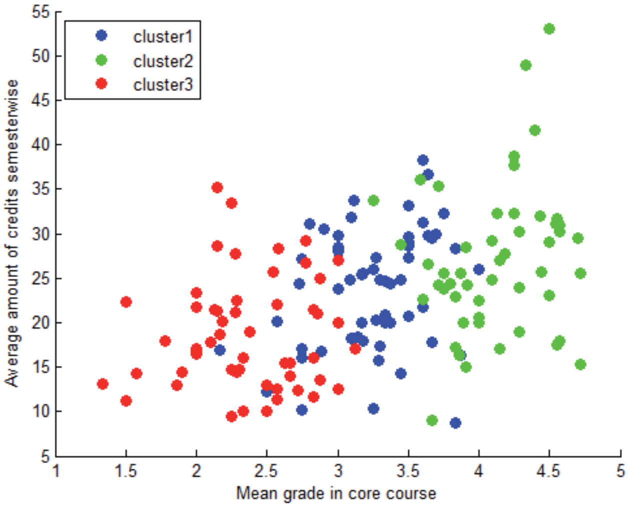


Figure 5: Semesterwise credits versus the mean grade in core courses for the students in each cluster.

was expected.

In terms of quality, the students in D_6 are clearly separated into the three clusters. To check whether the clusters also differentiate the students according to their quantity of studies, we looked also (see Section 2.1 and 3) at the students' average number of credits. In Figure 5, the semester-wise relation of grades in the core courses and the overall credits of the individual students in the different clusters is visualized. From this figure, we deduce that the students who belong to *cluster 2* not only are the best when it comes to the average grades in the core courses but also are the most efficient as they earn on average the most number of credits per semester. Likewise, the students in the gradewise low-performing *cluster 3* also earn the fewest credits per semester (on average eight credits less than the students in *cluster 2*). The correlation coefficient of the mean grade in the core courses and the average number of credits semesterwise per student is 0.4415 with a p-value that is highly statistically significant. We know that this relation does not exist in the whole student level and when the average of all studies is used (see Figure 1). Thus, we conclude that for the core CS courses, the students who perform well in terms of grades also perform well in terms of the number of courses.

5. Predictive analysis using multilayer perceptron

The goal, when addressing the third EDM category in this study, is to predict the mean grades and credits of the students given only the grades of the core courses they have passed. Similarly as in Section 4, we are interested in interpretable results, which here correspond to detecting the inputs (courses) that contribute to the prediction model the most. Concerning the model, multilayer perceptron (MLP) neural networks are universal nonlinear regression approximators (see, e.g., Pinkus 1999 and articles therein), which can be used in supervised learning. The feedforward MLP transformation starts directly from the input variables, different from other popular techniques such as radial basis function networks or support vector machines, which construct their basis in the space of observations. This is an appropriate starting point because our purpose is to assess the importance of the model inputs, which correspond to the core courses being analyzed. In this way, we close our between-method triangulation by contrasting the previous results and conclusions based on unsupervised analysis with the corresponding results from a supervised, predictive technique.

There are many inherent difficulties when a flexible model is used in prediction and trained using a given set of input-output samples. First, because of the universality, such a model could actually represent the discrete dataset precisely (e.g., Tamura and Tateishi 1997; Huang 2003), which would mean that all the noise in the samples would be reproduced. Thus, one needs to restrict the flexibility of such models. This can be done in two ways: by restricting the size of the network's configuration (number and size of layers; structural simplicity) or restricting the nonlinearity of the encoded function (size of weights, see Bartlett 1998; functional simplicity). Here we will assess the network's simplicity along both dimensions, in order to favor and restore the simplest model (cf. Occam's razor). Second, we look for a prediction model that provides the best generalization of the sample data, and, for this purpose, apply the well-known stratified cross-validation (see Kohavi 1995) to compute an estimate of the generalization error. Stratification means that, given a certain labeling to encode classes in a discrete dataset, the number of samples in the created folds (subsets) coincides with the sizes of the different classes as closely as possible. Clearly, the number of classes and number of folds do not need to be the same. Third, as in clustering, use of a local optimizer to solve the nonlinear optimization problem to determine the network weights provides only local search (exploitation), and for exploration, we use multiple restarts with random initialization (see Kärkkäinen 2002). The whole training approach as just summarized has been more thoroughly introduced and tested in Kärkkäinen (2014) and successfully applied in time-series analysis in Kärkkäinen et al. (2014).

Next we will derive and detail the whole predictive approach. First, the MLP neural network and its determination are formalized, and then the overall training algorithm and the input-sensitivity analysis are developed and described.

5.1. Prediction with input sensitivity analysis

5.1.1. MLP training approach

The action of the multilayer perceptron in a layered, compact form can be given by (e.g., Hagan and Menhaj 1994)

$$\mathbf{o}^0 = \mathbf{x}, \quad \mathbf{o}^l = \mathcal{F}^l(\mathbf{W}^l \tilde{\mathbf{o}}^{(l-1)}) \ \text{ for } l = 1, \dots, L. \tag{3}$$

Here the layer number (starting from zero for the input) has been placed as an upper index. By $\tilde{}$ we indicate the addition of bias terms to the transformation, which is realized by enlarging a vector $\mathbf{v}$ with constant: $\tilde{\mathbf{v}}^T = \begin{bmatrix} 1 & \mathbf{v}^T \end{bmatrix}$. In practice, this places the bias weights as the first columns of the layer matrices that then have the factorization $\mathbf{W}^l = \begin{bmatrix} \mathbf{W}_0^l & \mathbf{W}_1^l \end{bmatrix}$. $\mathcal{F}^l(\cdot)$ denotes the application of activation functions on the lth level. Formally, this corresponds to matrix-vector multiplication in which the matrix components are functions, and component multiplication is replaced with application of the corresponding component function (Kärkkäinen, 2002). The dimensions of the weight-matrices are given by $\dim(\mathbf{W}^l) = n_l \times (n_{l-1} + 1),\ l = 1, \dots, L,$ where n_0 is the length of an input-vector $\mathbf{x}$, n_L the length of the output-vector $\mathbf{o}^L$, and $n_l, 0 < l < L$, determine the sizes (number of neurons) of the hidden layers.

Using the given training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, with $\mathbf{x}_i \in \mathbb{R}^{n_0}$ denoting the input-vectors and $\mathbf{y}_i \in \mathbb{R}^{n_L}$ the output vectors, respectively, the unknown weight matrices $\{\mathbf{W}^l\}_{l=1}^L$ in (3) are determined as a solution of an optimization problem

$$\min_{\{\mathbf{W}^l\}_{l=1}^L} \mathcal{J}(\{\mathbf{W}^l\}). \tag{4}$$

We restrict ourselves to MLP with one hidden layer, and the actual cost function reads as follows:

$$\mathcal{J}(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2N} \sum_{i=1}^{N} \left\| \mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)(\mathbf{x}_i) - \mathbf{y}_i \right\|^2 + \frac{\beta}{2n_1} \sum_{(i,j)} \left(|\mathbf{W}^1_{i,j}|^2 + |(\mathbf{W}^2_1)_{i,j}|^2 \right) \tag{5}$$

for $\beta \geq 0$ and $\mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)(\mathbf{x}_i) = \mathbf{W}^2 \tilde{\mathcal{F}}^1(\mathbf{W}^1 \tilde{\mathbf{x}}_i)$. The special form of regularization omitting the bias column $\mathbf{W}^2_0$ is due to Corollary 1 by Kärkkäinen (2002): *Every locally optimal solution to* (4) *with the cost functional* (5) *provides an unbiased regression estimate having zero mean error over the training data*.

The universal approximation property guarantees the potential accuracy of an MLP network for given data and the unbiasedness as just described provides statistical support for its use, but as explained above, we also address the network's *simplicity* and *generalization*. Thus, in our actual training method we grid-search the size of the hidden layer n_1 and the size of the regularization coefficient β: The smaller n_1, the simpler the structure of the network; and the larger β, the smaller the weight values and the closer the MLP to a (simpler) linear, single-layered network. Moreover, cross-validation is used as the technique to ensure that generalization ability of the network is taken as the main accuracy criterion. Finally, the usual gradient-based optimization methods for minimizing (5) act locally, so that we repeat the optimization with random initialization twice when we search for the values of metaparameters n_1 and β. When they have been fixed, the final network is optimized using five local restarts to further improve the exploration of the search landscape.

The whole training approach for the MLP network is given in Algorithm 3. We use the following set of possible regularization parameter values, which were determined according to prior computational tests:

$$\vec{\beta} = \begin{bmatrix} 10^{-2} & 7.5 \cdot 10^{-3} & 5 \cdot 10^{-3} & 2.5 \cdot 10^{-3} & 10^{-3} & 7.5 \cdot 10^{-4} & 5 \cdot 10^{-4} & 2.5 \cdot 10^{-4} & 10^{-4} \end{bmatrix}.$$

The prediction error with a training or test set is computed as the mean Euclidian error

$$\frac{1}{N} \sum_{i=1}^{N} \left\| \mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)(\mathbf{x}_i) - \mathbf{y}_i \right\|. \tag{6}$$

We use the most common sigmoidal activation functions $s(x) = \frac{1}{1+\exp(-x)}$ for $\mathcal{F}^1$. All input variables are preprocessed into the range $[0, 1]$ of $s(x)$ to balance their scaling with each other and with the range of the overall MLP transformation (see Kärkkäinen 2002 for a more thorough argument).

5.1.2. Derivation of input sensitivity of MLP

To assess the relevancy of the input (see John et al. 1994; Kohavi and John 1997) of an MLP model, one basic technique is to estimate the sensitivity of the network's output compared to

```
Algorithm 3: Reliable determination of MLP neural network.
Input: Training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ .
Output: MLP neural network $\mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)$.
Define a vector $\vec{\beta}$ of regularization coefficients, maximum size of the hidden layer
$n1max$, and $nfolds$, the number of folds for cross-validation, created using stratified
random sampling;
for $n_1 \leftarrow 1$ to $n1max$ do
    for $regs \leftarrow 1$ to $|\vec{\beta}|$ (| · | denotes the size of a vector) do
        for $k \leftarrow 1$ to $nfolds$ do
            for $i \leftarrow 1$ to 2 do
                Initialize $(\mathbf{W}^1, \mathbf{W}^2)$ from the uniform distribution $\mathcal{U}([-1, 1])$;
                Minimize (5) with current $n_1$ and $\vec{\beta}(regs)$, and the CV Training set;
                Store Network for smallest Training_Set_Prediction_Error;
            end
            Compute Test_Set_Prediction_Error for the stored Network;
        end
        Store $n_1^* = n_1$ and $\beta^* = \beta$ for the smallest mean Test_Set_Prediction_Error;
    end
end
for $i \leftarrow 1$ to 5 do
    Initialize $(\mathbf{W}^1, \mathbf{W}^2)$ from $\mathcal{U}([-1, 1])$;
    Minimize (5) using $n_1^*, \beta^*$ and the whole training data;
end
```

its input. Seven possible definitions of sensitivity were compared in Gevrey et al. (2003) in an ecological context and four of them, further, in relation to chemical engineering in Shojaeefard et al. (2013). Both comparisons concluded that in order to assess the relevancy and rank the features, the partial derivatives (PaD) method proposed by Dimopoulos et al. (1995) provides appropriate information and computational coherency in the form of stability. Thus, we also use the analytic partial derivative as the core of the sensitivity measure, but in a more general and more robust fashion than Dimopoulos et al. (1995).

An analytical formula for the MLP input sensitivity can be directly calculated from the layer-wise formula (3). The precise result is stated in the next proposition.

Proposition 1

$$\nabla_{\mathbf{x}} \mathcal{N}(\{\mathbf{W}^l\})(\mathbf{x}) = \frac{\partial \mathbf{o}^L}{\partial \mathbf{x}} = \prod_{l=L}^{1} \operatorname{diag}\{(\mathcal{F}^l)'\}\, \mathbf{W}_1^l. \tag{7}$$

Here $\mathbf{W}_1^l$ denotes, as before, the lth weight matrix without the first bias column. In particular, for an MLP with one hidden layer and linear output ($\mathbf{o}^2 = \mathbf{W}^2\, \tilde{\mathcal{F}}^1(\mathbf{W}^1\, \tilde{\mathbf{x}})$), (7) states that

$$\frac{\partial \mathbf{o}^2}{\partial \mathbf{x}} = \mathbf{W}_1^2 \operatorname{diag}\{(\mathcal{F}^1)'\}\, \mathbf{W}_1^1. \tag{8}$$

```
Algorithm 4: Input sensitivity ranking.
  Input: Data $(\mathbf{X}, \mathbf{Y}) = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ of inputs and desired outputs.
  Output: Ranked list of MLP input variables.
  1: Fix $\vec{\beta}$ and $n1max$, and apply Algorithm 3 to obtain $\mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)$;
  2: Compute MAS of $\mathcal{N}(\mathbf{W}^1, \mathbf{W}^2)$ according to formula (9);
  3: Order input variables in descending order with respect to MAS to establish ranking;
```

With the discrete data $\{\mathbf{x}_i\}_{i=1}^N$, input sensitivity must be assessed and computed over the dataset. Thus, we apply (7) to compute the *mean absolute sensitivity, MAS* (see Ruck et al. 1990):

$$\frac{1}{N} \sum_{i=1}^{N} \left| \frac{\partial \mathbf{o}^L}{\partial \mathbf{x}_i} \right| \tag{9}$$

of the trained network for all input variables. After this formula is applied, the approach for input ranking is based on the following concept: The higher the MAS, the more salient the feature is for the network. This is due to the well-known Taylor theorem in calculus related to local approximation of smooth functions (see Apostol 1969). Namely, if a function is locally constant, its gradient vector (i.e., the vector of partial derivatives) is zero, and such a function could be (locally) represented and absorbed to the MLP bias. Thus, the larger the mean sum of the absolute values of the local partial derivatives for an input variable, the more important that input variable is for representing the variability of an unknown function approximated by the MLP. Thus, the descending order of MAS values defines the ranking of input variables over one run of Algorithm 3. The method described by Dimopoulos et al. (1995) starts with the similar analytic formula (formula (3)) as in (7), but (7) is a generalization because our MLP model contains the bias nodes in order to always guarantee unbiased regression estimate for the training data in Algorithm 3. Moreover, as with clustering, we compute the overall input-output sensitivity formulae using the robust *mean absolute error* instead of the sum-of-squares proposed in Dimopoulos et al. (1995), which nonuniformly concentrates on large deviations from zero (see Kärkkäinen and Heikkola 2004).

The whole algorithm for deriving the MLP input sensitivity is given in Algorithm 4. To this end, many points in this algorithm may produce variability in the final result, the ranking. With different runs, different foldings appear in cross-validation and different local initializations are tested when seeking the values of the metaparameters n_1 and β. Thus, it typically happens that a different final network is encountered from repetitions of Algorithm 3 whose ranking (1–12, where 1 represents the most significant) is then determined using Algorithm 4. To assess the stability and soundness of this result, we repeat Algorithm 4 five times, store the rankings obtained with different runs, and, then, compute the classical Fleiss kappa κ (Fleiss, 1971), which precisely quantifies the reliability of agreement between a fixed number of MLP network raters. The actual variable rating is then based on the ascending order of the sum of rankings from these five repetitions (between 5–60, where 5 means that such a variable was declared as the most significant for all the repetitions).

5.2. Predictive results and their analysis

As input data for MLP, we use the same set as in the cluster analysis, i.e. the grades of the students who have completed at least half of the core courses; see Table 3. Moreover, the missing values (29.65% altogether) are again completed by using the hot-deck imputation with 8 prototypes (see Section 4.3 for more thorough description). As output data, for each student considered, we use (i) the mean grade and (ii) the mean number of credits per semester, individually.

Results of the predictive analysis process, as described above, are provided in Table 8. There, for each course, the "RSum" provides the sum of rankings (1–12) of five individual runs of Algorithm 4. Moreover, in order to assess the stability of the final ranking, we have tested 3-fold, 7-fold, and 10-fold stratified cross-validation. As labels for the 3-fold stratification, we used the three cluster indices that were obtained in the previous section for $K = 3$ (the analyzed result). For the 7-fold CV, the labels corresponded to the number of completed courses in Table 3, i.e., to the separate groups of students for $q = 6, \ldots, 12$, whose sizes are given by n_q. In the third stratified cross-validation strategy with 10 folds, we used the labels that were obtained when clustering the students into 8 clusters (same as in imputation).

Thus, the strategy to create the different number of stratified folds was completely different, but the final rankings of the 7-fold and 10-fold CV were exactly the same, and there was only one very small difference compared to the 3-fold CV: For the mean grade, rankings of the *PCtools* and *Datanet* courses were swapped. We conclude that there is high reliability concerning the final rankings, because the Fleiss κ shows *moderate agreement* for grades with 7 and 10 folds and the rest of the cases witness *substantial agreement* between the ratings of the individual runs of Algorithm 4. From "MeanError" (see Table 8), which represents the mean of the prediction error (6) over the five runs, we conclude that mean grades can be predicted (in the generalization sense as explained above) about twice as accurately as the mean number of credits semesterwise. Again, this illustrates the higher and more random individual variability of the number of credits obtained per semester compared to the level of grades (see also Figures 6 and 7).

Based on the results presented in Table 8, we draw the following main conclusions: Compared to the correlation and clustering analysis results, also based on the predictive MLP input sensitivity analysis, the courses *Datanetworks* and *Computer Structure and Architecture* seem to be most influential to the overall performance in the studies. For the performance in grades, also the course *Object Oriented Analysis and Design* pops up, and, for the overall credits, the largest course *Programming 2* shows (as in the previous analyses) high significance.

For some course, like *Computer and Datanetworks as Tools* and *Programming of Graphical User Interfaces*, there is a large difference in the ranks between the mean grades and the mean credits, which was not addressed as strongly by the other two EDM techniques. One reason for this might be the varying number of students passing a course, which is reflected in the predictive analysis as the higher need of imputation. As can be seen from Figure 2, many fewer students have passed these two courses compared to the other courses[5].

The predictions and the prediction errors for grades and credits, studentwise, are illustrated in Figures 6 and 7. In the figures, the x-axis corresponds to a student index, where the students are taken in the ascending order for missing courses; the larger the index, the more core course grades are missing, and were imputed in the MLP training data. With this respect, the accuracy

[5] Actually, also fewer students passed *Research Methods in Computing*, but this course has already been found to be less influential and the most different to the courses (see especially the discussion in Section 4.3).

Table 8: Input rankings for the three foldings.

	3-fold CV				7-fold CV				10-fold CV			
	grades		credits		grades		credits		grades		credits	
MeanError	6.44e-3		1.22e-2		6.37e-3		1.22e-2		6.36e-3		1.22e-2	
Fleiss κ	0.76		0.72		0.49		0.62		0.52		0.78	
Course	RSum	*rank*	RSum	*rank*	RSum	*rank*	RSum	*rank*	RSum	*rank*	RSum	*rank*
PCtools	19	4	49	10	16	3	47	10	15	3	50	10
Datanet	17	3	10	2	19	4	10	2	18	4	10	2
OOA&D	10	2	39	7	12	2	40	7	13	2	39	7
Alg1	30	6	20	4	31	6	20	4	31	6	20	4
IntroSE	36	7	42	9	36	7	42	9	36	7	41	9
OpSys	39	8	57	11	41	8	57	11	41	8	57	11
DB&DMgm	45	9	15	3	42	9	15	3	42	9	15	3
Prog1	60	12	30	6	59	12	32	6	59	12	30	6
Prog2	50	10	5	1	50	10	5	1	50	10	5	1
CompArc	5	1	25	5	6	1	25	5	6	1	25	5
GUIprog	24	5	58	12	22	5	58	12	23	5	58	12
CompRes	55	11	40	8	56	11	41	8	56	11	40	8

of the mean number of credits per semester shows large increase at the end. As can be seen from Figure 6, the grades of the core courses predict, with reasonable accuracy, the overall mean grade level of a student. This result is promising, especially when the number of credits related to the analyzed core courses is typically less than half of the total number of credits; see Table 4.

In contrast, the generalization accuracy of the average number of credits per semester is very bad (see Figure 7), and the last students, i.e., those with the most missing values, are the most erroneous. Thus, we do not recommend the final network for actual prediction, but the network is considered suitable for the sensitivity analysis. The difference between accurate prediction and stable detection of input relevance is also clearly captured in Table 8 as explained above: The rankings in the repeated attempts in Table 8 are very stable, as shown by the Fleiss κ's, even if the prediction accuracy can be very poor as shown in Figures 6 and 7.

Hornik et al. (1989) summarize the essence of MLP training: "We have thus established that such 'mapping' networks are universal approximators. This implies that any lack of success in applications must arise from inadequate learning, insufficient numbers of hidden units or the lack of a deterministic relationship between input and target." The proposed training approach here tries to manage all these issues in order to end up with *the most reliably generalizing MLP network*. Thus, we try to capture the deterministic behavior within the data and use this to compute the input relevance. Stability of the results as witnessed in Table 8, with substantial within-method triangulation, supports the conclusion that this was obtained here.

6. Conclusions

This paper presents methods for detecting the main courses that determine the general success in CS-oriented studies. We employed techniques from the three main categories of educational data mining, partly working in relation to the remaining two categories as assistance in individual analyses. Moreover, we showed how to cope with the nonstructured sparsity pattern in data,

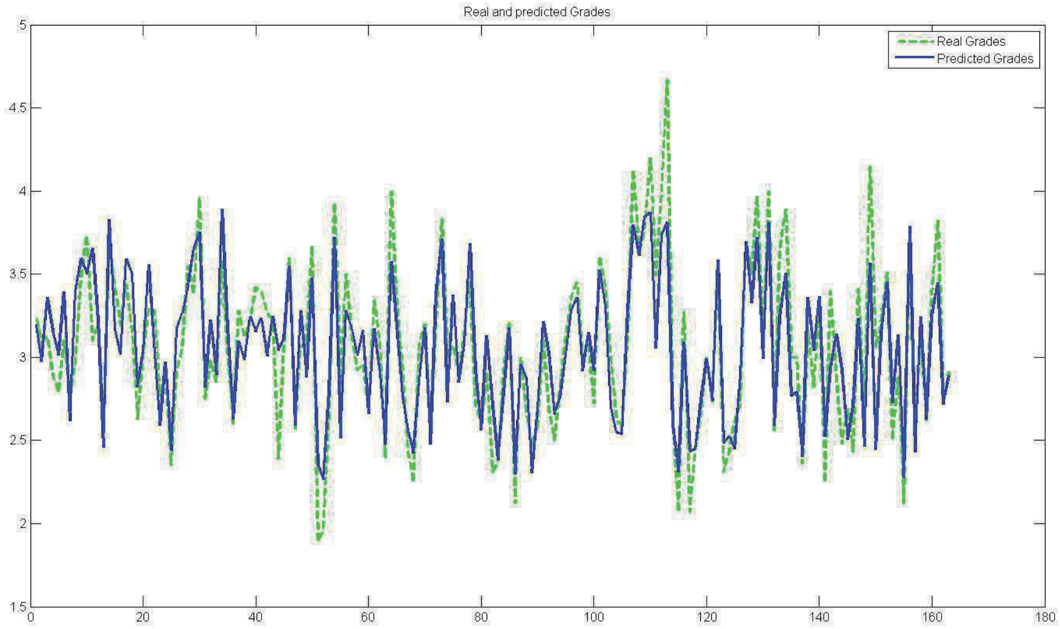


Figure 6: Prediction of mean grades: real (green) and predicted (blue) values.

using the available data strategy and prototype-based imputation. In Table 9, all analysis results are summarized. We can conclude the study from the educational domain level point of view and from the methodological point of view.

From the domain level point of view and based on Table 9, we conclude that the quality of studies is determined by the first introductory courses, *Datanetworks* and *Computer Structure and Architecture*, offered in the first year of the program. Both courses test more the general capability of a student to study than the actual knowledge of professional CS skills. Though they have technical topics, they are taught on a conceptual level, and especially compared with the third introductory course (see Table 2), they are completed by a final examination at the end of the course. Therefore, these courses test how well the student is able to learn, understand, and explain concepts instead of testing specific (IT) skills. When it comes to credits/timely graduation, a student's success is also determined by sedulousness and perseverance: The *Programming 2* which is also creditwise the largest course (see Table 2), is strongly related to the number of credits that a student can earn in general with hard work. Thus, for the overall performance, general study capabilities are more important than the occupational skills and students can succeed in CS studies with diligent and goal-oriented study behavior without being the most skilled programmers with mathematical talent. This is important knowledge that should be communicated to the students in the beginning of their studies.

Naturally, our conclusions from the organizational level as such are not generalizable to

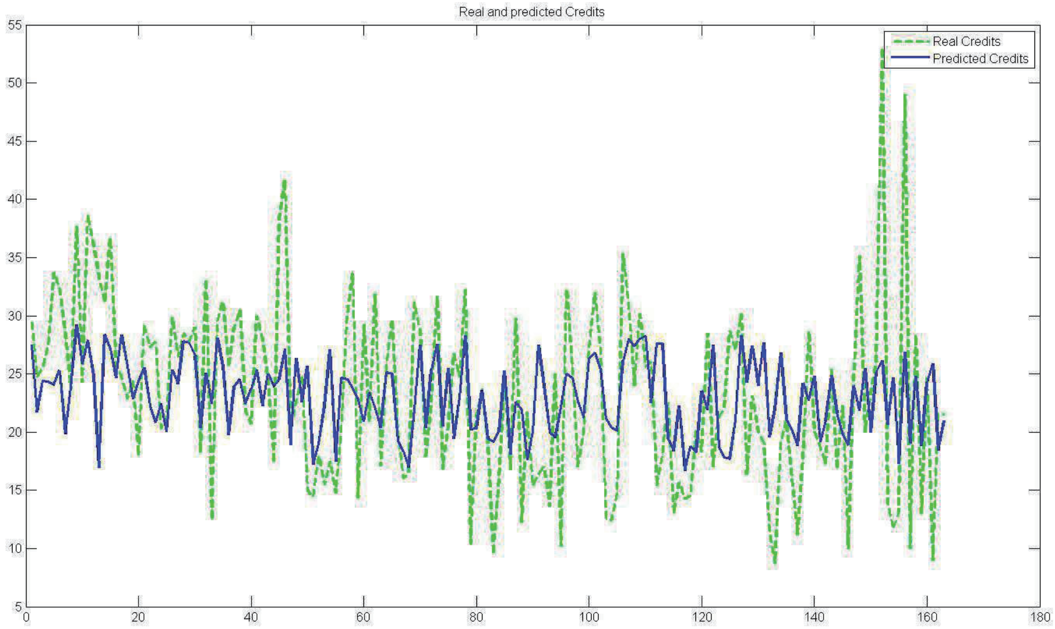


Figure 7: Prediction of mean credits: real (green) and predicted (blue) values.

other institutions since educational data and the subsequent knowledge of particular courses are different. From the methodological perspective, however, both overall approach and the individual methods with their varying but argumented details are general and can be applied to analyze the sparse data of student performance. If a snapshot of a study registry of an arbitrary educational institution were taken, there were missing values similarly to our case for the uncompleted courses. And, then, all methods and approaches could be applied. Furthermore, according to our current computational experience, we can conclude that for around a dozen variables (even using Matlab): i) correlation analysis scales up to one million observations, ii) clustering analysis scales up to hundreds of thousands of observations, and iii) predictive MLP analysis scales up to thousands of observations. This means that our methods can also be used for larger datasets.

In general, on the methodological level, the combination of within-method and between-method triangulation provided very solid results concerning the overall effects and impact of the analyzed courses. To deal with our student data, it was necessary to augment the existing methods and approaches to work with the sparse data. What about the soundness of the algorithms and the overall analysis presented here? There is lot of novelty in the procedures applied. The prototype-based clustering approach with available data spatial median as a statistical estimate is not a standard data mining technique. It was developed in the earlier work of the research group (Äyrämö, 2006; Kärkkäinen and Äyrämö, 2005), and its application is based on our own implementation throughout. Similarly, the way the clustering algorithm is construc-

Table 9: Summary of the results.

	grades				credits			
course	M1	M2	M3	sum (rank)	M1	M2	M3	sum (rank)
Computer and Datanetworks as Tools	11	9	3	23 (8)	12	9	10	31 (12)
Datanetworks	3	1	4	8 (2)	1	1	2	4 (1)
Object Oriented Analysis and Design	6	6	2	14 (4)	11	6	7	24 (7)
Algorithms 1	1	2	6	9 (3)	4	2	4	10 (3)
Introduction to Software Engineering	10	10	7	27 (10)	9	10	9	28 (10)
Operating Systems	8	7	8	23 (9)	8	7	11	26 (9)
Basics of Databases and Data Management	5	5	9	19 (6)	5	5	3	13 (5)
Programming 1	9	11	12	32 (11)	7	11	6	24 (6)
Programming 2	4	4	10	18 (5)	2	4	1	7 (2)
Computer Structure and Architecture	2	3	1	6 (1)	3	3	5	11 (4)
Programming of Graphical User Interfaces	7	8	5	20 (7)	6	8	12	26 (8)
Research Methods in Computing	12	12	11	35 (12)	10	12	8	30 (11)

tively initialized and how the variable ranking of prototypes is derived are not standard choices in cluster analysis. Moreover, the whole computational process for the predictive analysis — use of MLP with a) hot-deck imputation, b) complexity-aware training for best generalization, c) analytic formula-based robust input sensitivity derivation, d) sensitivity ranking, e) Fleiss κ as stability measure for rankings is completely novel. It is also based on our own implementation throughout. Training phase b) has been recently proposed and tested in Kärkkäinen (2014) and Kärkkäinen et al. (2014).

The underlying principle to study soundness in all the treatments here was based on local and global triangulation: In the correlation analysis, significancy was computed with and without Bonferroni correction. In cluster analysis, variable ranking was computed in two ways and assessed using the nonparametric Kruskal-Wallis test. Similarly, in the predictive analysis three different foldings (number of folds and how they are created) were used and Fleiss κ was then applied to the results of five iterations of the overall algorithm to study its stability. Thus, locally (for each method separately), we have made serious and versatile attempts to vary the meta-parametrization of the approaches and reported all the results. Globally, on the whole analysis level, we have again based our overall conclusions on the results and conclusions of the three methods of different orientations in EDM. We reason that such two-level treatment, where locally and globally the same results and their interpretation are supported by different approaches, improves the technical soundness of the study. Furthermore, the method for obtaining the final ranking, in clustering and in the MLP analysis, is novel and establishes a practical framework that can be used in similar applications.

Acknowledgement

The unknown reviewers are gratefully acknowledged; their suggestions and comments led to significant improvement in the original manuscript. Finally, we thank Sami Äyrämö for enlightening discussions on the clustering.

REFERENCES

ALDAHDOOH, R. T. AND ASHOUR, W. 2013. Dimk-means distance-based initialization method for k-means clustering algorithm. *International Journal of Intelligent Systems and Applications (IJISA) 5,* 2, 41.

APOSTOL, T. M. 1969. *Calculus, Volume 2: Multi-variable Calculus and Linear Algebra with Applications to Differential Equations and Probability*. Wiley.

ÄYRÄMÖ, S. 2006. *Knowledge Mining Using Robust Clustering*. Jyväskylä Studies in Computing, vol. 63. University of Jyväskylä.

BAI, L., LIANG, J., AND DANG, C. 2011. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems 24,* 6, 785–795.

BAI, L., LIANG, J., DANG, C., AND CAO, F. 2012. A cluster centers initialization method for clustering categorical data. *Expert Systems with Applications 39,* 9, 8022–8029.

BAKER, R. ET AL. 2010. Data mining for education. *International Encyclopedia of Education 7*, 112–118.

BAKER, R. S. AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining 1,* 1, 3–17.

BARTLETT, P. L. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *Information Theory, IEEE Transactions on 44,* 2, 525–536.

BATISTA, G. AND MONARD, M. C. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence 17*, 519–533.

BAYER, J., BYDZOVSKÁ, H., GÉRYK, J., OBŠIVAC, T., AND POPELINSKỲ, L. 2012. Predicting drop-out from social behaviour of students. In *Educational Data Mining 2012*. 103–109.

BHARDWAJ, B. AND PAL, S. 2011. Mining educational data to analyze students' performance. *(IJCSIS) International Journal of Computer Science and Information Security, 9,* 4.

BOUCHET, F., KINNEBREW, J. S., BISWAS, G., AND AZEVEDO, R. 2012. Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning. In *Educational Data Mining 2012*. 65–72.

BRADLEY, P. AND FAYYAD, U. 1998. Refining initial points for k-means clustering. In *ICML*. Vol. 98. 91–99.

BRYMAN, A. 2003. Triangulation. *The Sage encyclopedia of social science research methods. Thousand Oaks, CA: Sage*.

CALDERS, T. AND PECHENIZKIY, M. 2012. Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter 13,* 2, 3–6.

CAMPAGNI, R., MERLINI, D., AND SPRUGNOLI, R. 2012. Analyzing paths in a student database. In *Educational Data Mining 2012*. 208–209.

CARLSON, R., GENIN, K., RAU, M., AND SCHEINES, R. 2013. Student profiling from tutoring system log data: When do multiple graphical representations matter? In *Educational Data Mining 2013*. 12–20.

CHANDRA, E. AND NANDHINI, K. 2010. Knowledge mining from student data. *European Journal of Scientific Research 47,* 1, 156–163.

CHEN, L., CHEN, L., JIANG, Q., WANG, B., AND SHI, L. 2009. An initialization method for clustering high-dimensional data. In *Database Technology and Applications, 2009 First International Workshop on*. IEEE, 444–447.

CROUX, C., DEHON, C., AND YADINE, A. 2010. The k-step spatial sign covariance matrix. *Adv Data Anal Classif 4*, 137–150.

DENZIN, N. 1970. Strategies of multiple triangulation. *The research act in sociology: A theoretical introduction to sociological method*, 297–313.

DIMOPOULOS, Y., BOURRET, P., AND LEK, S. 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters 2,* 6, 1–4.

EMRE CELEBI, M., KINGRAVI, H. A., AND VELA, P. A. 2012. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*.

ERDOGAN, S. AND TYMOR, M. 2005. A data mining application in a student database. *Journal Of Aeronautics and Space Technologies 2*, 53–57.

FAYYAD, U., PIATESKY-SHAPIRO, G., AND P., S. 1996. Extracting useful knowledge from volumes of data. *Communications of the ACM 39,* 11, pp. 27–34.

FLEISS, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76,* 5, 378–382.

GEVREY, M., DIMOPAULOS, I., AND LEK, S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling 160*, 249–264.

HAGAN, M. T. AND MENHAJ, M. B. 1994. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Networks 5*, 989–993.

HALONEN, P. 2012. Tietotekniikan laitos. 2. TIETOTEKNIIKKA 12a- valintasyyt- opetuksen laatumielipiteet.pdf.

HAN, J., KAMBER, M., AND TUNG, A. 2001. Spatial clustering methods in data mining: A survey. *Data Mining and Knowledge Discovery*.

HARDEN, T. AND TERVO, M. 2012. Informaatioteknologian tiedekunta. 1. ITK 4- opinnoista suoriutuminen.pdf.

HARPSTEAD, E., MACLELLAN, C. J., KOEDINGER, K. R., ALEVEN, V., DOW, S. P., AND MYERS, B. A. 2013. Investigating the solution space of an open-ended educational game using conceptual feature extraction. In *Educational Data Mining 2013*. 51–59.

HAWKINS, W., HEFFERNAN, N., WANG, Y., AND BAKER, R. S. 2013. Extending the assistance model: Analyzing the use of assistance over time. In *Educational Data Mining 2013*. 59–67.

HETTMANSPERGER, T. P. AND MCKEAN, J. W. 1998. *Robust nonparametric statistical methods*. Edward Arnold, London.

HOLLANDER, M., WOLFE, D. A., AND CHICKEN, E. 2013. *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons.

HORNIK, K., STINCHCOMBE, M., AND WHITE, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 359–366.

HUANG, G. B. 2003. Learning capability and storage capacity of two-hidden-layer feedforward networks. *Neural Networks, IEEE Transactions on 14,* 2, 274–281.

HUBER, P. J. 1981. *Robust Statistics*. John Wiley & Sons Inc., New York.

JAIN, A. K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31,* 8, 651–666.

JERKINS, J. A., STENGER, C. L., STOVALL, J., AND JENKINS, J. T. 2013. Establishing the Impact of a Computer Science/Mathematics Anti-symbiotic Stereotype in CS Students. *Journal of Computing Sciences in Colleges 28,* 5 (May), 47–53.

JICK, T. D. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly 24,* 4, 602–611.

JOHN, G. H., KOHAVI, R., AND PFLEGER, K. 1994. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*. 121–129.

KÄRKKÄINEN, T. 2002. MLP in layer-wise form with applications in weight decay. *Neural Computation 14*, 1451–1480.

KÄRKKÄINEN, T. 2014. Feedforward Network - With or Without an Adaptive Hidden Layer. *IEEE Transactions on Neural Networks and Learning Systems*. In revision.

KÄRKKÄINEN, T. AND ÄYRÄMÖ, S. 2005. On computation of spatial median for robust data mining. *Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, EUROGEN, Munich*.

KÄRKKÄINEN, T. AND HEIKKOLA, E. 2004. Robust formulations for training multilayer perceptrons. *Neural Computation 16*, 837–862.

KÄRKKÄINEN, T., MASLOV, A., AND WARTIAINEN, P. 2014. Region of interest detection using MLP. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2014*. 213–218.

KÄRKKÄINEN, T. AND TOIVANEN, J. 2001. Building blocks for odd–even multigrid with applications to reduced systems. *Journal of computational and applied mathematics 131,* 1, 15–33.

KERR, D. AND CHUNG, G. 2012. Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining 4,* 1, 144–182.

KHAN, S. S. AND AHMAD, A. 2013. Cluster center initialization algorithm for k-modes clustering. *Expert Systems with Applications*.

KINNUNEN, P., MARTTILA-KONTIO, M., AND PESONEN, E. 2013. Getting to know computer science freshmen. In *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*. Koli Calling '13. ACM, New York, NY, USA, 59–66.

KOHAVI, R. 1995. Study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95)*. 1137–1143.

KOHAVI, R. AND JOHN, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence 97*, 273–324.

KOTSIANTIS, S. 2012. Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades. *Artificial Intelligence Review 37,* 4, 331–344.

MEILĂ, M. AND HECKERMAN, D. 1998. An experimental comparison of several clustering and initialization methods. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 386–395.

MENDEZ, G., BUSKIRK, T., LOHR, S., AND HAAG, S. 2008. Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education 97,* 1.

PINKUS, A. 1999. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 143–195.

RICE, W. R. 1989. Analyzing tables of statistical tests. *Evolution 43,* 1, 223–225.

ROUSSEEUW, P. J. AND LEROY, A. M. 1987. *Robust regression and outlier detection*. John Wiley & Sons Inc., New York.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika 63,* 3, 581–592.

RUBIN, D. B. AND LITTLE, R. J. 2002. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

RUCK, D. W., ROGERS, S. K., AND KABRISKY, M. 1990. Feature selection using a multilayer perceptron. *Neural Network Computing 2,* 2, 40–48.

SAARELA, M. AND KÄRKKÄINEN, T. 2014. Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Educational Data Mining 2014*. 60–68.

SAHAMI, M., DANYLUK, A., FINCHER, S., FISHER, K., GROSSMAN, D., HAWTHRONE, E., KATZ, R., LEBLANC, R., REED, D., ROACH, S., CUADROS-VARGAS, E., DODGE, R., KUMAR, A., ROBINSON, B., SEKER, R., AND THOMPSON, A. 2013a. Computer science curricula 2013.

SAHAMI, M., ROACH, S., CUADROS-VARGAS, E., AND LEBLANC, R. 2013b. ACM/IEEE-CS Computer Science Curriculum 2013: Reviewing the Ironman Report. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*. ACM, New York, USA, 13–14.

SAN PEDRO, M. O. Z., BAKER, R. S., BOWERS, A. J., AND HEFFERNAN, N. T. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Educational Data Mining 2013*. 177–184.

SHOJAEEFARD, M. H., AKBARI, M., TAHANI, M., AND FARHANI, F. 2013. Sensitivity analysis of the artificial neural network outputs in friction stir lap joining of aluminum to brass. *Advances in Material Science and Engineering 2013*, 1–7.

SPRINGER, A., JOHNSON, M., EAGLE, M., AND BARNES, T. 2013. Using sequential pattern mining to increase graph comprehension in intelligent tutoring system student data. In *Proceeding of the 44th ACM technical symposium on Computer science education*. ACM, 732–732.

STEINBACH, M., ERTÖZ, L., AND KUMAR, V. 2004. The challenges of clustering high dimensional data. In *New Directions in Statistical Physics*. Springer, 273–309.

TAMURA, S. AND TATEISHI, M. 1997. Capabilities of a four-layered feedforward neural network: Four layers versus three. *IEEE Transactions on Neural Networks 8,* 2, 251–255.

VALSAMIDIS, S., KONTOGIANNIS, S., KAZANIDIS, I., THEODOSIOU, T., AND KARAKOS, A. 2012. A clustering methodology of web log data for learning management systems. *Educational Technology & Society 15,* 2, 154–167.

VIHAVAINEN, A., LUUKKAINEN, M., AND KURHILA, J. 2013. Using students' programming behavior to predict success in an introductory mathematics course. In *Educational Data Mining 2013*. 300–303.

XU, R. AND WUNSCH, D. C. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks 16,* 3, 645–678.

ZHONG, C., MIAO, D., WANG, R., AND ZHOU, X. 2008. Divfrp: An automatic divisive hierarchical clustering method based on the furthest reference points. *Pattern Recognition Letters 29,* 16, 2067–2077.

PII

KNOWLEDGE DISCOVERY FROM THE PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT

by

Mirka Saarela, Tommi Kärkkäinen 2017

Book chapter in Learning Analytics: Fundaments, Applications, and Trends.
Springer, pp. 229–267

Chapter 10
Knowledge Discovery from the Programme for International Student Assessment

Mirka Saarela[1]**, Tommi Kärkkäinen**[2]

[1] University of Jyväskylä, Department of Mathematical Information Technology
40014 Jyväskylä, Finland
mirka.saarela@jyu.fi

[2] University of Jyväskylä, Department of Mathematical Information Technology
40014 Jyväskylä, Finland
tommi.karkkainen@jyu.fi

Abstract. The Programme for International Student Assessment (PISA) is a worldwide study that assesses the proficiencies of 15-year-old students in reading, mathematics, and science every three years. Despite the high quality and open availability of PISA data sets, which call for big data learning analytics, academic research using this rich and carefully collected data is surprisingly sparse. Our research acts on this deficit by discovering novel knowledge from PISA through the development and use of appropriate methods. Since Finland has been the country of most international interest in the PISA assessment, a relevant review of the Finnish educational system is provided. This chapter also gives a background on learning analytics and presents findings from a novel case study. Similarly to the existing literature on learning analytics, the empirical part is based on a student model, but differently from these, our model represents a profile of a national student population. We compare Finland to other countries by hierarchically clustering these student profiles from all the countries that participated in the latest assessment, validating the results through statistical testing. Finally, an evaluation and interpretation of the variables that explain the differences between the students in Finland and those of the remaining PISA countries is presented. Based on our analysis, we conclude that, globally, learning time and good student-teacher relations are not as important as collaborative skills and unassumingness explaining students' success in the PISA test.

Keywords: PISA, Learning Analytics, Big Data, Knowledge Discovery, Hierarchical Clustering

Abbreviations

ESCS	Economic, social and cultural status
LA	Learning Analytics
OECD	Organisation for Economic CO-operation and Development
PISA	Programme for International Student Assessment

10.1 Introduction

The original purpose of *Learning Analytics* (LA), as stated, e.g., by Siemens (2013, p. 1383) and Ferguson (2012, p. 306), was to "measure, collect, analyze, and report data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs." Slightly different variants were later given to characterize the discipline (Pardo & Teasley 2014, Gray et al. 2014, Siemens & Baker 2012). Increased attention to Massive Open Online Courses (e.g., Wang et al. 2014, Ye & Biswas 2014, Reich et al. 2014, Coffrin et al. 2014, Hickey et al. 2014, Santos et al. 2014, Vogelsang and Ruppertz 2015, Ferguson and Clow 2015, Hansen and Reich 2015, Wise et al. 2016, Hecking et al. 2016) has widened the need for data based learning support in the direction of the so-called *big data*. This is evidenced by several articles (e.g., Picciano 2012, Chatti et al. 2012, Siemens 2012, Chatti et al. 2014, Dawson et al. 2014, Wise & Shaffer 2015, Merceron et al. 2016) and also by the theme of the 2015 Learning Analytics and Knowledge conference "Scaling up: Big data to Big Impact" (see Dawson et al. 2015).

PISA is a worldwide triennial survey conducted by the Organisation for Economic Cooperation and Development (OECD), resulting in publicly available educational data on a large scale. Besides assessing the proficiency of 15-year-old students from different countries and economies in reading, mathematics, and science, PISA provides "data about learners and their contexts" being one of the largest public databases[1] of students' demographic and contextual data, such as their attitudes and behaviors towards various aspects of education. More than seventy countries and economies have already participated in PISA, and the assessment is referred to as the "world's premier yardstick for evaluating the quality, equity and efficiency of school systems" (OECD 2013*a*).

In the PISA studies, data collection is of very high quality, including the development of the data collection instruments, the procedures of data collection, and the storage of the data in public databases. This is evidenced by the large amount of money spent on ensuring quality related to these issues. However, much less money has been invested in the analysis of the collected data and only few PISA analysis studies have resulted in publications in the scientific field (Olsen 2005*a*). Rutkowski et al. (2010) argue that the sizes of PISA data sets as well as the technical complexities within them may be the reason why more researchers do not work with these freely available and high quality data.

Our research is motivated by the lack of secondary analysis of PISA data, which calls for the development and utilization of big data LA methods for making discoveries within the international domain of PISA. Such methods can then be used to summarize the PISA data sets in novel ways in order to better understand students from diverse countries and the settings in which they learn (Siemens & Baker 2012). Hence, in relation to big data LA, we focus on the interna-

[1] PISA data can be downloaded from http://www.oecd.org/pisa/pisaproducts/.

tional context, trying to understand national education systems as learning environments. Such a scope of LA was also emphasized by Long & Siemens (2011), who pointed out that LA should occur on the national and international levels, primarily targeting for national governments and education authorities. Similarly as a classroom is in a school is in a city is in a region is in a country is in a continent, thorough use of educational data and empirical evidence should be linked to principles and practices of educational systems that are known to have an effect on learning. This is the primary concern in PISA.

Chatti et al. (2014) introduced a reference model for LA based on four dimensions: stakeholders, objectives, data, and methods, resembling the critical LA dimensions suggested by Greller & Drachsler (2012). Fig. 10.1 illustrates how large-scale educational assessments, such as PISA, can leverage big data LA according to these dimensions. Namely, national bodies introduce the objectives (i.e., factors that constitute good national education systems) for assessing the international student population. Then, large data representing student background and proficiency are sampled and transformed into derived representations, whose characteristics (sample to population alignment introducing weights, rotated test design introducing missing values) must be handled by the applied LA methods. When meaningful patterns are found, they are reported back to the educational decision makers.

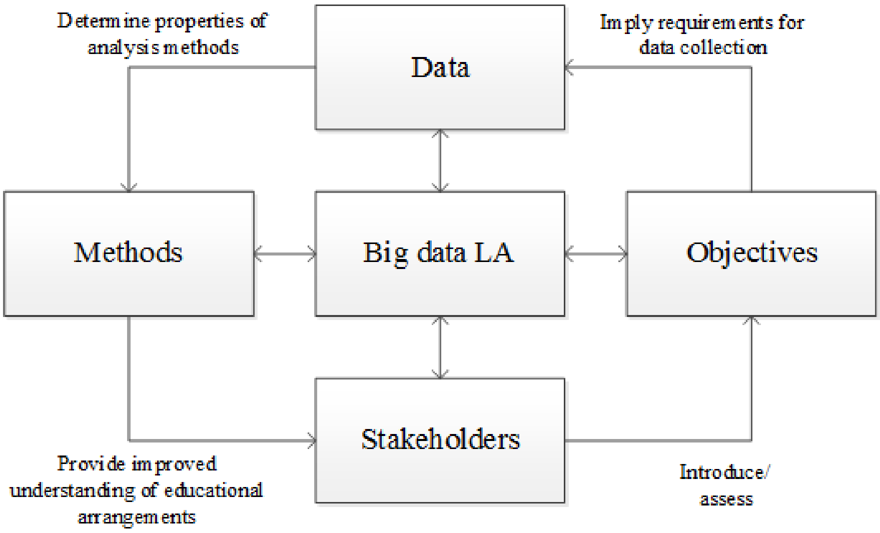

Determine properties of analysis methods
Data
Imply requirements for data collection
Methods
Big data LA
Objectives
Stakeholders
Provide improved understanding of educational arrangements
Introduce/ assess

Fig. 10.1 Conducting big data LA for large-scale educational system assessments (cf. Chatti et al. 2014, Greller & Drachsler 2012).

Ferguson et al. (2014) emphasize the large-scale institutional adoption of appropriate educational patterns. In the best case, the institutional meso-level approaches are aggregated from the upscale local micro-level patterns and from the downscale macro-level characteristics of a good educational system. Thus, meaningful patterns at the macro-level, e.g., within a large educational organization, originate from characteristics of a large student population in relation to the rigorously measured learning outcomes.

The structure of this paper is as follows. In Section 10.2, we provide necessary background on big data LA and educational knowledge discovery from PISA. In Section, 10.3, a relevant review on methodologically related studies is provided and the forms and complexities of PISA data are described. Afterwards, the overall analysis method is depicted in Section 10.4. In Section 10.5, the results and interpretations of the hierarchical clustering of the aggregated country profiles are presented and statistically validated. In Section 10.6 the PISA results are visualized in a dashboard. Finally, in Section 10.7, the empirical work is summarized and in Section 10.8, the overall conclusions are given.

10.2 Background and Related Work

We provide next the necessary theoretical background for the empirical part of the chapter. First, we explain big data LA and summarize LA methods. Then, we characterize a pool of methodologically related work on the use of clustering in educational data analysis. We observe that methodologically related studies are typically conducted on the micro-level of individual courses or tutoring systems.

10.2.1 Towards Big Data LA

As emphasized in the introduction, LA studies are increasingly leveraging big data. The term “big” in big data does not solely refer to the amount of data, but actually refers to four ‘V’s (the first three according to Laney (2001), the last one as described, e.g., by Gupta et al. (2014)): (i) *Volume* refers to the size of data sets caused by the number of data points, their dimensionality, or both; (ii) *Velocity* is linked to the speed of data accumulation; (iii) *Variety* stands for heterogeneous data formats, which are caused by distributed data sources, highly varying data gathering, etc.; and (iv) *Veracity* refers to the fact that (secondary) data quality can vary significantly, and manual curation is typically impossible.

In relation to the big data LA, PISA data are characterized by high volume and low veracity due to missing values, but no velocity and small, well-managed variety due to the meticulous design. Moreover, unlike the existing LA studies, the collected student sample is aligned to the whole worldwide student population under study through the use of the weights (see the last paragraphs in Section 10.2.3). For example, the sample data of PISA 2012 consists of circa half a million students, representing 24 million 15-year-old students from 68 different countries and territories.

Chatti et al. (2012) state that different LA techniques for detecting interesting educational patterns originate from four analysis categories: *statistics*; *information visualization*; *data mining* (identifying this with knowledge discovery in databases) in the form of classification, clustering, and association rule mining; and *social network analysis*. The other LA researchers support this notion that data mining and knowledge discovery techniques are one category of the broad LA methods. Rogers (2015), for example, lists data mining as one of the more sophisticated quantitative methods in LA and Siemens (2013) states that the knowledge discovery from databases is an LA technique that became increasingly important.

Generally, with the advent of big data in education, the LA methods have shifted from the more traditional data analysis techniques, such as statistics, to the more scalable data mining methods (Hershkovitz et al, 2016; Joksimović et al, 2016). In fact, Ferguson (2012) points out that the two main differences between the general educational research and the specific research field of LA according to the LA definition given in the beginning of this chapter is that LA "make use of pre-existing, machine-readable data, and that its techniques can be used to handle 'big data.'"

Application of data mining and knowledge discovery methods in an educational context typically realizes an *educational knowledge discovery process* that, especially when using an open educational dataset like PISA, supports learning and knowledge analytics (Verbert et al 2012). Several case studies (e.g., Hu et al, 2106; Brown et al, 2016; Grawemeyer et al, 2016; Allen et al, 2016; Chandra & Nandhini 2010) have proved the need and success of specific knowledge discovery processes and data analysis methods within the educational domain. However, data from many of the existing educational case studies are specific for certain educational environments or institutions, which complicate the comparison of the techniques and the provided results.

In contrast, PISA tests are standardized, and the resulting data sets are comparable between different nations and their educational arrangements. Hence, PISA provides an interesting and novel case for big data LA techniques (Saarela & Kärkkäinen 2014, Saarela & Kärkkäinen 2015*a,b,c*, Kärkkäinen & Saarela 2015), combining methodological requirements, due to the above-mentioned technical complexities of the data, with comparative, educational knowledge discovery.

10.2.2 On Educational Data Analysis Using Clustering

As pointed out above, *clustering* is one of the key techniques in the data mining category of the LA methods. Next we describe a pool of work related to clustering of educational data as well as the empirical work in Sections 10.4-10.5. This set of papers was mostly identified by scanning through the most relevant publication forums (see Saarela et al. 2016*a*) in the field, especially the *Journal of Learning An-*

alytics[2] and the *Conference on Learning Analytics & Knowledge*[3], restricting to the topic of clustering with some real educational data set. The description of the work is organized according to the used clustering method and the size of the clustered educational data set.

Hierarchical Clustering. Logs of 454 online mathematics practice sessions by 69 students were clustered by Desmarais & Lemieux (2013). In that study, the preprocessing first transformed the logs into temporal sequences (time series) reflecting the state of interaction between the student and the learning environment. These representations were then clustered using an agglomerative hierarchical method, and the interpretation of the result was based on visualizing the clusters as state sequence diagrams. Three characteristic forms of using the system were identified: (i) exploratory browsing, (ii) short practice sessions, and (iii) exercise intensive sessions.

Self-regulatory strategies of undergraduate students, especially characteristics of accessing online learning material, were studied by Colthorpe et al. (2015). Hierarchical clustering of 97 students was able to separate high and low performing students, where at first sight extensive use of lecturing recordings indicated poorer academic performance. This could, however, be explained by the form of engagement in the learning material.

Segedy et al. (2015) provided more in-depth analysis of student's self-regulated interaction with the learning material in an open-ended computer-based learning environment. Student assessment was based on the coherence analysis, whose descriptive metrics for 99 sixth grade students were, as part of the versatile analysis process, separated into five clusters using complete-link hierarchical clustering. In addition to two very small clusters of (i) confused guessers and (ii) students disengaged from the task, the main clusters characterized the self-regulated interaction patterns of (iii) frequent researchers and careful editors, (iv) strategic experimenters, and (v) engaged and efficient students.

Hu et al. (2016) used hierarchical clustering to analyze responses of 523 English and Chinese primary school students to a questionnaire about their reading behaviors, preferences and attitudes towards reading. Three main reading profiles were identified and they were fully characterized by good, moderate, and bad reading habits.

Hecking et al. (2016) combined social similarity (i.e., distances in the communication graph of the students) and semantic similarity (i.e., distances between the content-based roles by the students) to construct a socio-semantic blockmodelling approach for analyzing a MOOC discussion forum. Hierarchical clustering was used in the actual construction of the blockmodel from the derived similarity measure. The analysis of the communication graph of 647 students in 502 threads on 27 forums verified the presence of different roles, with moderate correlation

[2] See http://learning-analytics.info/ .

[3] See http://lakXX.solaresearch.org/, where XX stands for year in which the conference took place. For example, http://lak16.solaresearch.org/ contains a link to the proceedings of the 2016 conference.

between a social and a semantic role by a student. Discovery of the three main socio-semantic roles suggested that online discussion forums need better recognition and adaptation to the different user roles.

K-Means. A collaboration of 31 participants in a math discussion board, through the lens of activity theory, which links individual and social behavior, was addressed by Xing et al. (2014) using the prototype-based k-means clustering method. In this study, the important phases of the educational clustering process, preprocessing and interpretation of the clustering result, were strongly present. The result consisted of three clusters characterizing (i) personally participative but on the group level less communicative learners, (ii) collaboratively participating but shallow learners, and (iii) less participative poor learners.

An automated approach using the k-means clustering algorithm was described by Li et al. (2013), for constructing a student model from the content features of algebra problems. Methodologically versatile preprocessing (feature extraction, min-max scaling, principal component analysis) and ten-fold cross-validation characterized the approach. The experiment with data from 71 students concluded that the clustering-based model was at least as good as the prior manually constructed model, being able to reveal previously unidentified and valuable knowledge components of mathematical problem solving. An innovative assessment of the physical learning environment also using the k-means clustering method was reported by Almeda et al. (2014). The result consisted of four different clusters characterizing the similar content profiles of 30 classroom walls, as decorated by the teachers.

Multiple clustering methods (including k-means and hierarchical clustering) at various stages of the data analysis were applied by Blikstein et al (2014), to reveal the different patterns and trends of the development of programming behavior in an introductory undergraduate programming course. The overall analysis of 370 participants and 154,000 code snapshots was concluded in multiple ways. Firstly, for different tasks within LA one needs different kind of tools ranging from fast and simple wrap-ups of data into advanced machine learning methods running on high-performance computing platforms. Secondly, concerning the clustering methods, one needs either better support to interpret the result of a clustering method or application of more advanced methods to improve the insight and knowledge discovery from data. Thirdly, concerning the domain of the study, the changes in the code update patterns by the students were more strongly correlated with the course performance compared to the size of code updates.

A subset of methods used by Blikstein et al. (2014) were also utilized by Worsley & Blikstein (2014) to analyze the problem-solving patterns of 13 students for open-ended engineering tasks. The LA method was based on segmentation and extraction of action features from the hand-coded video data. The k-means algorithm produced four clusters, whose interpretation could be summarized into two principal dimensions of idea quality and design process, which were both related to students' level of experience.

Expectation-Maximization. Derived variables of multiple thematic groups from the log data of 106 college students using an intelligent tutoring system fostering self-regulated learning, was clustered by Bouchet et al. (2013). They used the expectation-maximization algorithm from Weka resulting in three clusters as suggested by the knee point (see Saarela & Kärkkäinen 2015*a*), after careful cross-validation with multiple restarts. The three clusters were mostly characterized by the varying levels of performance, but also reflected (through metadata) differences in the number of self-regulated learning processes in which the students were engaged. Bogarin et al. (2014) also used the expectation-maximization algorithm from Weka and discovered three clusters from the log data of 84 Psychology students learning to learn online with Moodle. Especially a cluster of the most passive online students was detected, of which two-thirds failed the course.

Activity in online discussion forums, as a predictor of study success, was also studied by López et al. (2012). Methodologically it was shown that the prototypes obtained from the expectation-maximization clustering algorithm with ten-fold cross-validation with Weka software were able to distinguish 114 different and informative cases of university student behavior. Similarly to Bogarin et al. (2014), it was concluded that active participation in the course forum was a good predictor of the final mark for the course.

Summary. To summarize this small survey on educational clustering, hierarchical clustering, k-means, and expectation-maximization were the most common approaches. This was also concluded in the review by Peña-Ayala (2014). Similarly, student modeling, including behavior and performance models, was the dominant educational data analysis approach, covering all except Almeda et al. (2014) of the assessed research (see Table 11 in the work published by Peña-Ayala 2014). Note that a set of older references concerning the use of clustering in educational settings, as briefly introduced by Bouchet et al. (2013) in Section 6, also emphasized the student model as an important part of intelligent, online tutoring systems.

10.2.3 Learning Analytics Approaches Oriented to Analyze PISA Repositories

As concluded in the previous section, clustering is one of the key techniques for analyzing educational data, especially in LA. However, most of the educational clustering studies use small data sets of tens or at most hundreds of students at the micro- and meso-level of educational systems. By comparison, the PISA 2012 data set is comprised of around half a million students and represent a population of 24 million people worldwide (see the last paragraph in Section 10.3.2).

A considerable amount of literature has been published on PISA. However, as observed by Olsen (2005*a*), these publications are mainly national or international reports that have not gone through the peer-review process. Furthermore, many of the peer-reviewed publications dealing with PISA (e.g., Deng & Gopinathan 2016, Auld & Morris 2016, Rasmussen & Bayer 2014, Yates 2013, Bank 2012, Bulle 2011, Waldow et al. 2014, Grek 2009, Simola 2005, Sahlberg 2011, Kumpulainen & Lankinen 2012) do not perform own empirical analysis, but only refer to the reports or statistics published by the OECD. In the papers where own empirical models are being derived and analyzed (e.g. Skryabin et al. 2015, Kriegbaum et al. 2015, Erdogdu & Erdogdu 2015, Tømte & Hatlevik 2011, Zhong 2011, Fonseca et al. 2011) the missing data is mostly completely removed and only the sample is analyzed by ignoring the weights and, hence, the population level. Moreover, usually students from only a few countries are being compared in the existing literature, although a very scarce pool of whole PISA sample level comparisons exists (e.g., Drabowicz 2014, Zhong 2011).

We have also carefully assessed the use of clustering with PISA data sets and have only been able to identify our own recent publications for PISA 2012 (Saarela & Kärkkäinen 2014, Saarela & Kärkkäinen 2015*b*,*c*) and one older publication for PISA 2003 (Olsen, 2005*b*). Thus, our main contributions here are that we augment the traditional PISA analysis by utilizing big data LA methods and by working with the whole data on the macro-level of the whole student population, confining to the recommendations given by the OECD (2014*b*). This population level scope is a novel setting in big data LA.

10.3 PISA Profile

In this section, we outline contextually related work of the chapter. More precisely, since Finland is of main interest in our clustering application, we introduce the main characteristics of the Finnish educational system that has performed so well in the PISA assessments as well as related research. The last part of this section is devoted to a description of the collection and overall processing of the PISA assessment, yielding to multiple forms of publicly available educational data sets on a macro-level.

10.3.1 The Finnish Educational System and PISA

In this paper, our main focus is on Finland in comparison to the other countries that participated in the latest PISA assessment. Traditionally, Finnish students have performed exceptionally well in the PISA tests. The reasons for Finland’s success in PISA, particularly in the 2003 and 2006 assessment cycles, have been analyzed in several studies and educational stakeholders from all over the world have visited Finland to find explanations for the high performing students.

Consequently, education became an important asset in Finland’s image and identity. In fact, Finland has invested considerably in the international educational export sector (Schatz et al. 2016), and although Finland’s place in the international ranking dropped in the latest PISA assessment, it is still placed the highest in Europe. Here, our goal is to assess the variables that most distinguish Finland from the other countries participating in PISA.

Finland's high performance in the PISA assessments has been analyzed in several articles. Many of these articles have linked the well-performing students to the highly qualified teachers, who need to have a Master's degree for a permanent position. In particular, it has been argued that in Finland, being a teacher is one of the most prestigious occupations, as evidenced by the fact that only the best and most motivated students are admitted to the teacher training programs, as well as the observation that Finnish teachers enjoy a very high status in the society (Morgan 2014, Sahlberg 2011, Linnakylä et al. 2011, OECD 2011, Andere 2015).

A second reason that has been identified to contribute to Finland's high results in PISA relates to the organization of the national school system. Instead of (i) market-oriented schooling, (ii) standardization of schools and tests, concentrating on measurable performance, and (iii) competition between students and schools, the focus in Finland's schools is more on cooperation, collaboration, and the belief that teachers will support each student's individual learning (Simola 2005, Sahlberg 2011). National curricula as well as explicit learning objectives and standards do exist, but schools and teachers in Finland enjoy great autonomy and decision-making authority, i.e., they can decide on learning strategies and pedagogical methods in order to reach the common educational goals (Kumpulainen & Lankinen 2012, Linnakylä et al. 2011, OECD 2011).

The fact that schools in Finland are neither competing nor are evaluated by standardized tests is one of the reasons why the variance between the Finnish schools is so small[4] (Simola 2005). Additionally, there is a no division of students into different school types or tracks based on their performance. Indeed, all students in Finland attend common, untracked, comprehensive schools of equally good quality from grades 1–9, typically those nearest to their homes. These schools are publicly funded and offer free lunches, healthcare, and school transport for all pupils (OECD 2011, Linnakylä et al. 2011).

These mutually interdependent and interconnected factors that are associated with Finland's high achievements in PISA have also been emphasized by Välijärvi et al. (2007) who have concluded that Finland's success can be explained by a combination of "comprehensive pedagogy, students' own interests and leisure activities, the structure of the education system, teacher education, school practices and, in the end, Finnish culture" (see Table 10.1).

Research has shown that culture tends to affect both people's goals and their actions to reach these goals (Hitlin & Piliavin 2004). As already pointed out above, Finnish people put great emphasis on equity and equality. Several studies have also highlighted the trust that seems to exist in Finnish culture in general, and between the educators and the community in particular (Sahlberg 2011, OECD 2011).

[4] According to the 2012 assessment, the between school variation in Finland is only 6% of the overall math performance which is the second lowest figure in comparison with all PISA countries.

Table 10.1 Interaction between culture and education in Finland.

Culture	Education
strong mutual trust	parents and government trust teachers (indicated by strong autonomy and authority of the teachers)
equity & equality (care for others instead of wanting to be the best)	common untracked comprehensive school systems, free lunch, health care and school transport, children with special needs study in the same classroom
indulgent country	minimal time allocated to studying, broad rich curriculum

The *Hofstede Model* (Hofstede 2011) acknowledges the idea of Finland being more a collaborative than a competitive country. According to the model, Finland’s society can be characterized as being highly “feminine,” meaning that the most important driving factors in life are to live a good life and to care for others instead of focusing on one’s own success and wanting to be the best. This is interesting when linked to the recent study by French et al. (2015), who found a negative causal relationship between education expenditure and power distance and masculinity. According to this study, the less masculine a country is the more it invests in education.

10.3.2 Characteristics and Forms of the PISA Data

According to the OECD, PISA results have a high degree of validity and reliability (for example, OECD 2014*b*, 2012), so they can be used to assess and compare the educational systems of the participating countries. To ensure the validity and reliability of PISA data large amounts of money are spent. For example, in Germany alone, the aggregate costs of PISA assessment have reached 21.5 million euros (Musik 2016). However, as already pointed out in the introduction of this chapter, the PISA assessments as well as the resulting PISA data are methodologically very complex.

As highlighted by the OECD (2012), “the successful implementation of PISA depends on the use, and sometimes further development, of state-of-the-art methodologies and technologies”. Since a mixture of different methods is used in this large study and many variables are derived, it is not obvious how certain values in the publicly available database[5] (see Fig. 10.2) were collected, obtained, and reported. The fact that PISA data are not trivial can also be concluded based on the time that is needed to publish the PISA data and results: Usually around 1.5 years

[5] Can be downloaded from http://pisa2012.acer.edu.au/downloads.php

passes after data collection before the first PISA results and data are published. For example, the 2012 PISA data collection took place in spring 2012, and its results were published in December 2013.

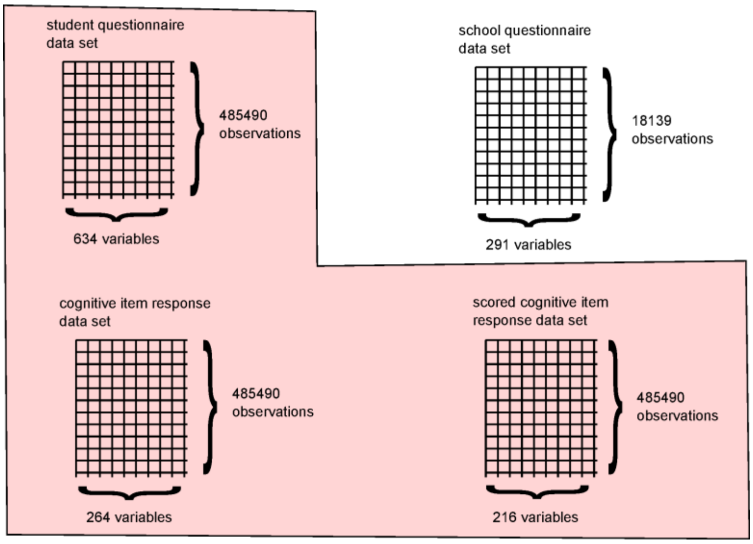

student questionnaire data set
485490 observations
634 variables
school questionnaire data set
18139 observations
291 variables
cognitive item response data set
485490 observations
264 variables
scored cognitive item response data set
485490 observations
216 variables

Fig. 10.2 Overview of the 2012 data sets available from OECD.

An overview of the 2012 PISA data is provided in Fig. 10.2. In all three data sets with pink backgrounds in Fig. 10.2, the assessed students are the observations. The basic information about the student (student's ID, country, test language, and school ID), and which test he or she was administered (booklet ID) is provided in all three of these student data sets. The *student cognitive items* and *scored cognitive item response* data sets document the students' responses to the cognitive items and how these were scored. Altogether there were 206 different cognitive items in the PISA 2012 data. An example of a cognitive item variable label is "SCIE—P2006 Wild Oat Grass Q4." As can be seen, it includes the domain (in this case, science), the PISA cycle in which the question was used first (in this case, PISA 2006), the name for the particular task unit[6] (in this case, *Wild Oat Grass*), and the question number (in this case, *4*).

The most informative and meaningful part of PISA data is the *student questionnaire data set* (see Fig. 10.2). However, as pointed out before, one of the biggest challenges when working with PISA data is that many variables in this data set are not direct measurements but rather already transformed and preprocessed variables. For example, the students' abilities/performances in the cognitive tests are summarized in the form of so-called *plausible values*. Plausible values are, as Wu (2005) puts it, "multiple imputations of the unobservable latent achievement for each student." This is explained more thoroughly at the end of this section.

[6] PISA items are organized into units. Each unit consists of a stimulus (consisting of a piece of text or related texts, pictures, or graphs) followed by one or more questions.

Certain scale indices in the data—indicating, for example, student's attitudes towards school and learning—are also derived variables. This means that in order to be able to work with PISA data, one need to understand how the many derived variables have been created and how they can be used for further analysis. In PISA, the *Rasch model*, which is a special case of item response theory, is used for this purpose.

Gray et al. (2014) emphasize the importance of integrating item response theory factors and methods, such as the Rasch model, to the existing LA models. Item response theory models can improve existing models because they can model latent (i.e., not directly measurable) traits, such as intelligence, ability or motivation. Moreover, they can be applied even with a large number of missing values. The potential of using item response theory in LA has been shown, for example, by Bergner et al. (2015) who estimated student abilities based on homework scores from a massive open online course of which a large number of scores were missing.

The second challenge when working with PISA data is the high sparsity. Since the assessment material developed for PISA exceeds the time that is allocated for the test, each student is administered only a fraction of the whole cognitive testing material and only one of the three different background questionnaires. Because of this rotated design, very few variables in PISA data sets have values for all observations. For example, in PISA 2012, each student was assigned a test booklet of cognitive items that should be solvable in two hours. However, the whole PISA 2012 cognitive item battery consisted of test items to be solved in six hours.

The *scored item set* (see Fig. 10.2) incorporates 206 scored items for 485,490 students. Nevertheless, because of the different booklets, which always contain only a fraction of the total items, 74% (that is, 738,604,20) of the different item variables have missing values. Similarly, because of the three different background questionnaires administered, the majority of the variables in the *student questionnaire data set* are missing approximately one-third of their values. We have discussed sparsity in educational data, particularly in PISA data, and algorithms to cope with this issue in many of our recent studies (Saarela & Kärkkäinen 2014, Saarela & Kärkkäinen 2015*a,b,c*, Kärkkäinen & Saarela 2015, Saarela et al. 2016*b*).

Finally, PISA data are an important example of large data sets that include weights. Only a fraction of 15-year-old students from each country take part in the assessment, but multiplied with their respective weights, which simply measure how many similar students one student in the sample represents, the gathered sample depicts the whole student population. For example, the sample data of the latest assessment consist of 485,490 students, which taking the weights into account, are representative of more than 24 million 15-year-old students in the 68 different countries and territories that participated in PISA 2012.

Both over- and under-sampling has taken place in PISA for different student groups. As a consequence, in order to state findings that are valid for the whole population, it is important to utilize the weights at each stage of the analysis. The way in which we incorporated the weights into a robust clustering algorithm for sparse data is illustrated and applied in our prior works (respectively, Saarela & Kärkkäinen 2015*c,b*).

10.3.3 Rasch Model

As described above, because of the different PISA test booklets administered, the actual scored student test data is extremely sparse with a great deal of missing values (74%). The easiest approach for measuring each student's ability would be to average the percentage of the correct answers over the three domains. However, since not all students obtained the same test items and as the test items varied in their difficulty, this approach is considered unreliable. With the Rasch model, however, the probability of a success on an item can be modeled as a logistic function of the difference between the student and item parameters (Rasch 1960). Hence, the Rasch model enables a comparison of student abilities/test results/characteristics, even if not all students were tested on the same test items.

In PISA, the Rasch model is employed both to estimate student abilities—depending on their item responses and the item difficulties in the cognitive test—and to estimate general student characteristics—depending on their responses on the background questionnaire. Mathematically, in the simplest case of the Rasch model when the test item is dichotomous, the probability that a student i with ability denoted by β_i provides a correct answer to an item j of difficulty δ_j can be stated as follows (10.1):

$$P\left(X_{ij} \middle| \beta_i, \delta_j\right) = \frac{\exp(\beta_i - \delta_j)}{1+\exp(\beta_i - \delta_j)}. \tag{10.1}$$

When the Rasch model is employed, it iteratively creates a continuum/scale on which both a student's ability and item difficulty are located, and where a probabilistic function links these two components. Usually, the item difficulties are estimated first, and this is referred to as the item calibration. The overall objective is to obtain data that will fit the model.

A student should give a correct answer to an easy item with higher probability than to a difficult item. Similarly, a student with high ability should give correct answers to items with higher probability than a student with low ability. This is shown in Fig. 10.3, where the probability that a correct answer is given to an item with difficulty δ = 0.6 is plotted for different student abilities. Moreover, as also illustrated in Fig. 10.3, when a student's ability is equal to the difficulty of the item, there is by definition a 50% chance of a correct response in the Rasch model.

To estimate the item difficulty, only the probability of being correct on that item and the ability of the students who completed the item must be known. Likewise, to estimate the student's ability, only the probability of being correct on a set of items and the difficulty of those items must be known (Embretson & Reise 2013). Every item and every student will be located in the created scale with the Rasch model. Therefore, comparable student ability estimates can be obtained, even if the students were assessed with a different subset of items (OECD 2014*b*). The only requirement is that some link items exist (i.e., some items in the different test booklets must be the same).

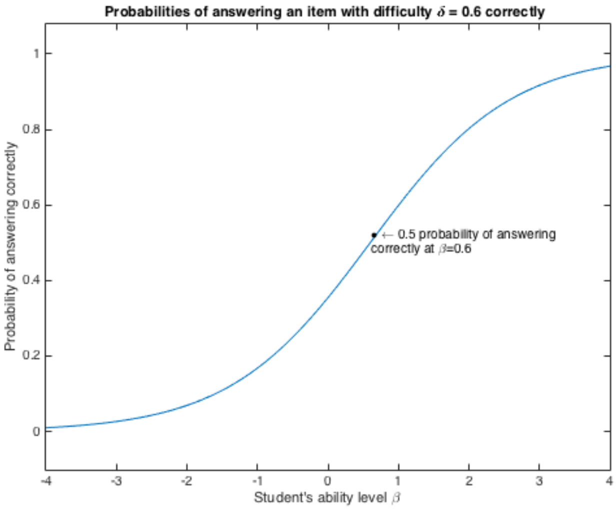


Fig. 10.3. Rasch model example. Probabilities that a correct answer is given to an item with difficulty δ = 0.6 for different student abilities. The probability that a student with ability β = 0.6 will provide a correct answer to this item is 0.5.

In PISA, a generalization of the original Rasch model is employed that can score not only dichotomous but also polytomous items (e.g. cognitive items can be scaled as *incorrect*, *partially correct*, and *correct* or questionnaire Likert-scale data can be scaled as *completely agree*, *agree*, *neutral*, *disagree*, and *completely disagree*). This model is called the one-parameter logistic model for polytomous items.

10.3.4 Plausible Values

There exist many other international, large-scale educational assessment studies such as PISA, including the *National Assessment of Educational Progress*[7], the *European Survey on Language Competences*[8], the *Trends in International Mathematics and Science Study*, and the *Progress in International Reading Literacy Study*[9]. The idea in PISA and in these other assessments is not to measure and report proficiencies of individual students. Instead, the primary goal is to provide a reliable overview of the proficiencies and national characteristics of the whole population (OECD 2014*b*, Marsman 2014). This is the main difference between typical micro- or meso-level LA and the big data LA for PISA.

[7] nces.ed.gov/nationsreportcard/

[8] www.surveylang.org/

[9] See both http://timssandpirls.bc.edu/

Plausible values are used to estimate the proficiencies of the population, which in PISA are all 15-year-old pupils within the participating countries. Some studies (Monseur & Adams 2008, Wu & Adams 2002, OECD 2014*b*) have shown that plausible values—in comparison to Weighted Likelihood Estimates, which overestimate, and Expected A Posteriori estimators, which underestimate population variances—produce unbiased estimates for population statistics.

In short, plausible values are random draws from the posterior distribution of a student's ability. These posterior distributions are estimated with a Bayesian approach in combination with the Rasch model. Hereby, the posterior distribution of a student's ability β_i, given his or her vector of item responses $\boldsymbol{x}_i$ and certain additional variables about the student from the background questionnaire (e.g. gender and many others) that are encoded in a vector $\boldsymbol{y}_i$, is defined as (10.2):

$$f(\beta\,|\boldsymbol{x}_i, \boldsymbol{y}_i) \propto P(\boldsymbol{x}_i|\beta, \delta) f(\beta\,|\boldsymbol{\lambda}, \boldsymbol{y}_i), \tag{10.2}$$

where $P(\boldsymbol{x}_i|\beta, \delta)$ denotes a Rasch model given the student's ability β and the difficulties of the items δ in the test, and $f(\beta\,|\boldsymbol{\lambda}, \boldsymbol{y}_i)$ denotes a population model. This population model for a student i is usually estimated with the latent (called latent because the predictor is unobserved) regression model $\beta_i = \boldsymbol{y}_i^T \boldsymbol{\lambda} + \varepsilon_i$, where $\varepsilon_i = \mathcal{N}(0, \sigma^2)$ (Marsman 2014, OECD 2014*b*).

In other words, in each country, the student abilities are assumed to follow a conditional Gaussian distribution, given $\boldsymbol{y}_i$, i.e., the variables from the background questionnaire. This is the prior distribution. Then, the student takes the PISA test. The statistical model ("likelihood") of the success in the test is a Rasch model, where the probability of success is a logistic function of the unknown but estimated latent ability and the difficulties of the test items (see Equation 10.1 and 10.2).

The estimated posterior distribution of the ability of the student is specific for each student, as each student has different values of background variables and test results. This means that success in the PISA test "corrects" our prior beliefs regarding the student's ability. If a student successfully solves a difficult item, this indicates higher ability than success on an easy item. However, the student's exact ability is not known, and is represented on the population level with five plausible values that are a random realization based on his or her posterior distribution. For this reason, when analyzing student performance, the official PISA protocol (OECD 2012) requires that the same analysis be repeated five times, separately for each plausible value.

10.4 Comparison of Students in PISA 2012 Countries Using Aggregated, Hierarchical Clustering

The empirical part of this work is focused on comparing Finland (through the student characteristics) to the other countries that participated in the PISA assessment 2012. This comparison is conducted by utilizing three of the four LA techniques described by Chatti (2012) (see Section 10.2.1): clustering as one of the core *data mining* techniques, *visualization* of the clustering result to illustrate Finland's position in comparison to the other countries, and, finally, *statistical* testing to verify the findings.

10.4.1 Variables for the Clustering

Our overall analysis method here is to apply hierarchical clustering on all PISA 2012 countries/economies, to visualize the similarities between the participating countries through a dendrogram, and to conduct different statistical tests on two different levels. For this, we first aggregated the entire half a million student sample of PISA 2012 into the population level of each country by computing the weighted means of the available data in a country-wise manner. We used all observations in the PISA 2012 data set. All variables in the PISA student data set (and their possible values) can be found in the codebook[10]. In Saarela & Kärkkäinen (2014), Saarela & Kärkkäinen (2015*c,b*) and Kärkkäinen & Saarela (2015), we have utilized the individual variables on a student level that are known to explain performance in mathematics. Here, we used an extended set of variables, including also those that are more on a classroom- (e.g. teacher behavior) or country- (e.g., time of formal instruction in a certain school subjects) than on an individual student level.

In Table 10.2, all variables used in this study are listed. All these variables are derived variables, constructed with the Rasch model using students' answers of the background questionnaire or other already derived variables. For example, the first variable, the *index of economic, social and cultural status*, is constructed using the *highest parental occupation*, the student's *home possessions*, and the *highest parental education*, which themselves are derived variables constructed with the Rasch model (OECD 2014*b*).

The following five variables, i.e., those with the IDs 2-6 in Table 10.2, are generally associated with performance on a student level, while the next ten variables (IDs 7-16) are all related to attitudes towards mathematics. Since mathematics was the major domain in 2012, attitudes towards this subject received considerable attention in the background questionnaire. Here, we use all ten mathematics indices that together summarize 67 items in the student background questionnaire.

The next five variables in the table (IDs 17-21) are related to how much time students spend studying. Both, formal learning time in different subject areas as well as out-of-school study hours are detailed. The last variable, *Age at ISCED 1* reports the beginning of the systematic apprenticeship of reading, writing, and mathematics. The last six variables (IDs 22-27) are all on a teacher or teaching level.

[10] Available at http://pisa2012.acer.edu.au/downloads/M_stu_codebook.pdf.

Table 10.2 Overview and identification of the derived PISA variables utilized in this study.

PISA variable	ID	PISA variable	ID
Economic, social and cultural status	1		
Sense of belonging	2	Attitude towards school: learning outcome	3
Attitude towards school: learning activities	4	Perseverance	5
Openness to problem solving	6		
Self-responsibility for failing in math	7	Interest in mathematics	8
Instrumental motivation to learn math	9	Self-efficacy in mathematics	10
Anxiety towards mathematics	11	Self-concept in mathematics	12
Behaviour in mathematics	13	Intentions to use mathematics	14
Subjective norms in mathematics	15	Mathematics Work Ethic	16
Out-of-School Study Time	17	Learning time (min. per week) - Test Language	18
Learning time (min. per week) – Mathematics	19	Learning time (min. per week) – Science	20
Age at <ISCED 1>	21		
Teacher Student Relations	22	Mathematics Teacher's Support	23
Teacher Behaviour: Formative Assessment	24	Teacher Behaviour: Student Orientation	25
Teacher Behaviour: Teacher-directed Instruction	26	Experience with Applied Math Tasks at School	27

10.4.2 Hierarchical Clustering

As pointed out above, a high number of values are missing in the PISA data. Moreover, each student in the PISA data sets has a weight expressing how representative he or she is for the population of all 15-year-old students within his or her country. Therefore, we computed for each country/economy the weighted means of the available data for each variable as inputs for the clustering algorithm. We then normalized our data set using z-scoring and applied hierarchical clustering with Matlab's default settings, i.e., agglomerative single-linkage clustering with the Euclidean distance.

Agglomerative clustering techniques operate in a bottom-up fashion (Zaki & Meira Jr 2014). Hence, we started with each PISA country as a separate cluster. Then, the most similar country clusters C_m and C_n were repeatedly merged so that they formed a new bigger cluster. The most similar clusters were defined as the ones with the smallest Euclidean distance between a point in C_m and a point in C_n (10.3):

$$\delta(C_m, C_n) = \min\{\delta(\boldsymbol{u}, \boldsymbol{v}) | \boldsymbol{u} \in C_m, \boldsymbol{v} \in C_n\}, \qquad (10.3)$$

where $\delta(u, v) = \left(\sum_{i=1}^{d}(u_i - v_i)^2\right)^{\frac{1}{2}}$ (see Zaki and Meira Jr 2014).

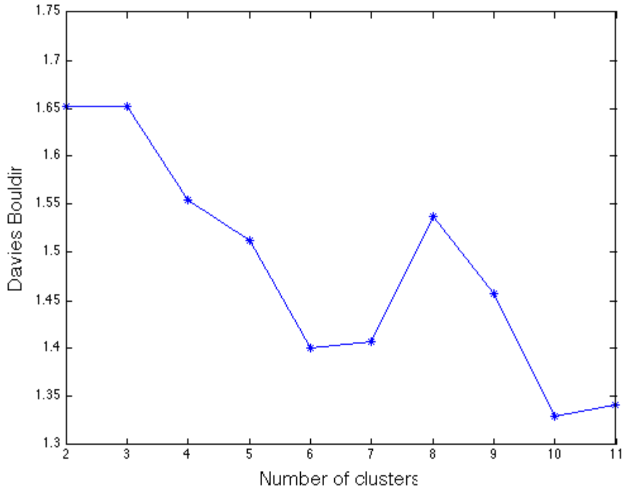


Fig. 10.4 The Davies-Bouldin index suggest that there are ten clusters in the data.

To decide the number of clusters in PISA 2012, the Davies–Bouldin cluster index (Davies & Bouldin 1979) was applied on the z-scored data. As can be seen from Fig. 10.4, the Davies-Bouldin index suggested that there are ten clusters in the data. Therefore, the merging of closest clusters was terminated after ten clusters were formed.

10.5 Results

In this section, we first present the clustering result and profile that. Then, the clustering results are analyzed more deeply using statistical tests on two different levels.

10.5.1 Visualization and Profiling of the Clusters

Fig. 10.5 shows the hierarchical clustering result. Based on the similarities of countries in particular groups, we suggest the following labels for the ten clusters as documented in Table 10.3.

Table 10.3 Clustering results.

ID	label	countries
C1	'Nordic/ English-speaking'	Australia, Canada, United-Kingdom, New-Zealand, Florida (USA), Connecticut (USA), Massachusetts (USA), USA, Denmark, Iceland, Norway, Sweden
C2	-	Costa-Rica, Israel, Uruguay
C3	'Eastern countries'	Bulgaria, Lithuania, Montenegro, Perm-Russia, Romania, Russia, Serbia
C4	'South America/ Africa'	Argentina, Chile, Tunisia
C5	'developing countries'	Brazil, Colombia, Indonesia, Mexico, Malaysia, Peru, Thailand, Turkey, Vietnam
C6	'high performing Asian'	Shanghai-China, Singapore
C7	'Kazakhstan'	Kazakhstan
C8	'Arabic'	UAE, Jordan, Qatar
C9	'Asian'	Hong-Kong-China, Japan, Korea, Macao-China, Taiwan
C10	'Europe'	Austria, Belgium, Switzerland, Czech-Republic, Germany, Spain, Estonia, Finland, France, Greece, Croatia, Hungary, Ireland, Italy, Liechtenstein, Luxembourg, Latvia, Netherlands, Poland, Slovak-Republic, Slovenia

It is a surprise that Finland is not part of the Nordic/English-speaking cluster, to which all other Nordic countries belong. This finding is interesting compared to the classification of Bulle (2011), who introduces "the Northern model: Denmark, Finland, Iceland, Norway, Sweden" as one of the five main OECD educational systems. Hence, if the educational systems are similar, this does not mean that the student characteristics are also similar.

The dendrogram implies that Finland belongs to the Europe cluster and is actually closest to the Netherlands. In the PISA 2012 results summary (OECD 2014*a*, page 7), the performances of these two countries in mathematics, among many other pairs of countries, were found to not be statistically significantly different. In addition, both the Netherlands and Finland are, according to the *Hofstede Model* (Hofstede 2011), highly feminine cultures.

As explained above, it was unexpected that Finland belonged to the Europe cluster and not to the Nordic/English-speaking cluster. To assess the significance of the single variables and to explain why a particular country was allocated to a certain cluster, we utilized statistical tests. Since not all of our variables were normally distributed, we had to use non-parametric tests.

To specifically address the finding of Finland's position, we will first report the differences between all the clusters. Second, we will summarize the differences between Finland and its own Europe cluster, and third, we will describe the variables that separate the Europe cluster from the Nordic/English-speaking cluster.

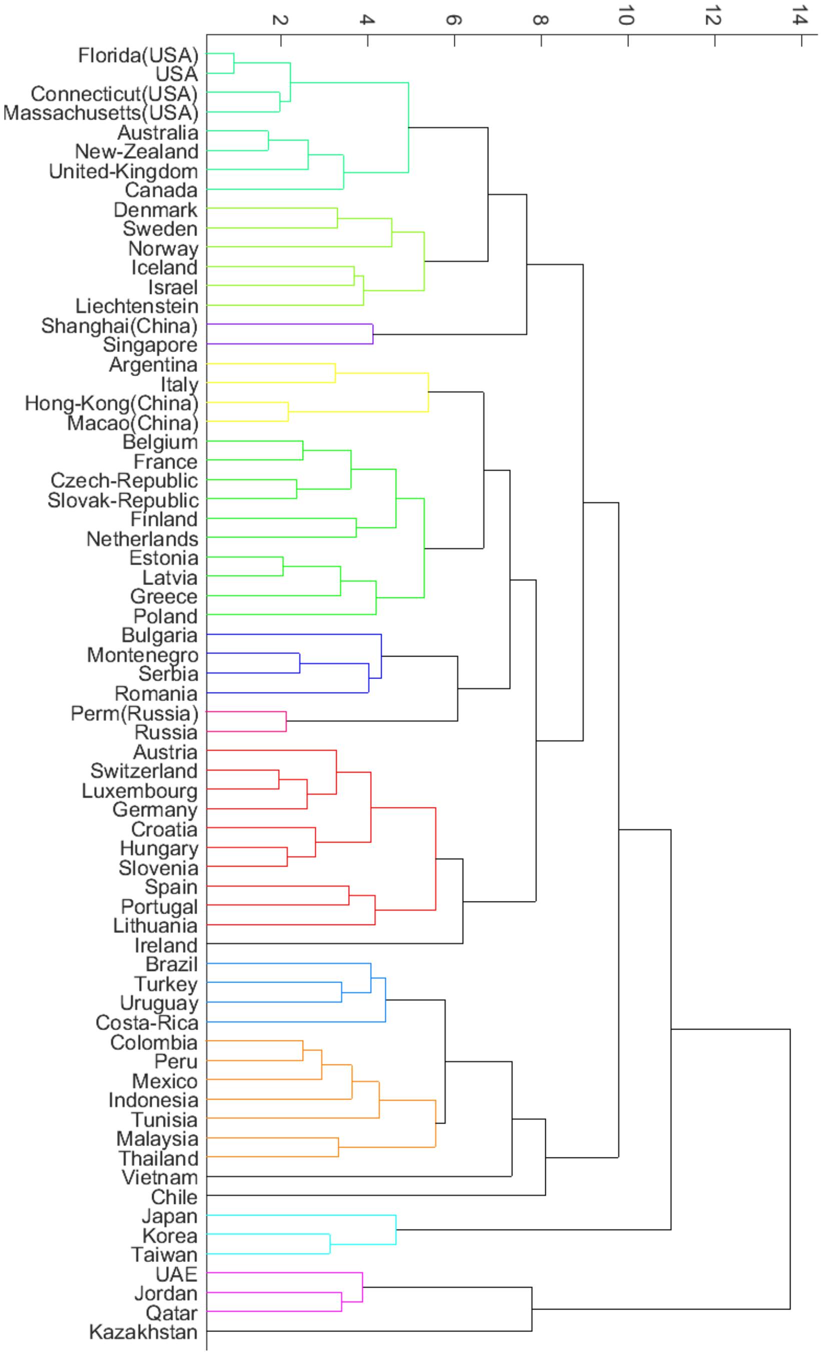


Fig. 10.5 Dendrogram of all countries when their weighted mean is clustered.

10.5.2 Differences between All the Global Clusters

A Kruskal-Wallis H test (Kruskal & Wallis 1952) showed that there was a highly statistically significant difference in 20 of the 27 variables between the different clusters. The test statistics of all highly statistically significant variables are provided in Table 10.4. With reference to Table 10.4, the variable 25 *teacher behaviour: student orientation*, i.e., how much attention teachers pay to individual students, was the most important in terms of accounting for variance in the cluster membership (χ^2(9) = 51,227, $p < 0.001$).

Table 10.4 Kruskal-Wallis H test statistics (all clusters) with a post hoc test.

variable	χ^2(9)	p	Post hoc test	variable		χ^2(9)	p	Post hoc test
1	48,676	★★★	C10-C5, C1-C5	4		38,499	★★★	C9-C1
5	33,306	★★★	-	7		37,399	★★★	-
8	48,701	★★★	C10-C5	9		49,857	★★★	C9-C5, C10-C5
10	30,765	★★★	-	11		42,170	★★★	C1-C5
12	35,298	★★★	-	13		49,549	★★★	C1-C5
14	34,029	★★★	-	15		49,082	★★★	C10-C5
16	39,863	★★★	-	18		40,457	★★★	-
19	36,542	★★★	-	22		42,940	★★★	C10-C5
23	46,378	★★★	C10-C5	24		45,203	★★★	-
25	51,227	★★★	C10-C5	26		42,610	★★★	-

Subsequently, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed highly statistically significant differences in the *ESCS* between the developing (mean rank = 5.67) and the Nordic/English-speaking cluster (mean rank = 57.25) as well as between the developing and the Europe (mean rank = 40.47) cluster, but not between any other group combination for this variable. This is also illustrated in Fig. 10.6, in which all pairwise comparisons of the different clusters for their ESCS are shown. In the figure, black lines reflect a pairwise comparison that is not statistically significant, while orange lines reflect a statistically significant pairwise comparison.

The last column in Table 10.4 summarizes the post hoc analysis for all the variables. As can be seen from the table, highly statistically significant differences were found in the *attitude towards school: learning activities*, i.e., the degree to which a student sees hard work in school pay off later, between the Asian (mean rank = 5.00) and the Nordic/English-speaking cluster (mean rank = 51.08), in the *interest in* and *enjoyment of mathematics* between the developing countries (mean rank = 56.89) and Europe (mean rank = 14.90) cluster, in the *instrumental motivation to learn mathematics*, i.e., the degree to which a student's hard work in mathematics pays off later, between the developing (mean rank = 57.89) and the Asian (mean rank = 7.80) countries, and between the developing countries and the Europe (mean rank = 19.10) cluster.

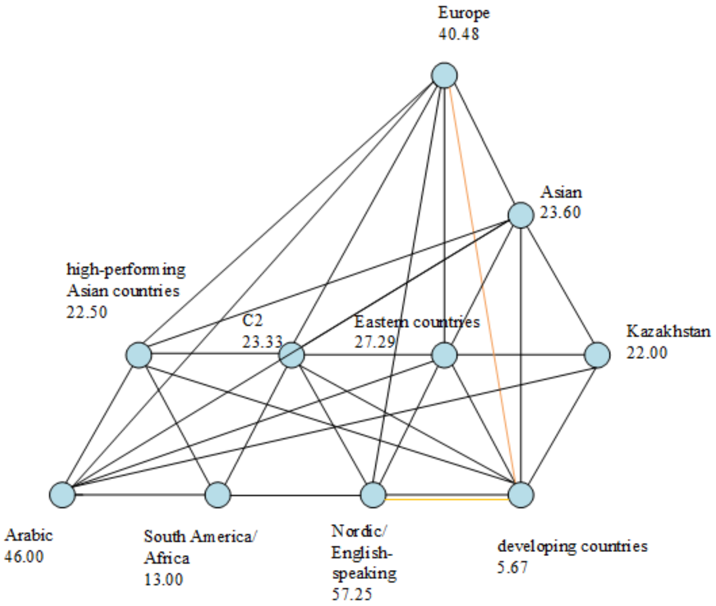


Fig. 10.6 Pairwise comparisons of clusters for ESCS. Statistical significant differences (developing countries-Nordic/English-speaking and developing countries-Europe) are marked yellow.

The developing countries cluster was revealed to be highly statistically significant different from the Nordic/English-speaking cluster with regard to the *anxiety towards mathematics* (mean rank C5 = 55.00 vs. C1 = 14.92) and the *behaviour in mathematics*, i.e., the role of mathematics inside and outside school, (mean rank C5 = 54.11 vs. C1 = 12.17). In addition, the developing countries cluster was found to be highly statistically significant different from the European cluster with regard to the *subjective norms in mathematics* (mean rank C5 = 51.11 vs. C10 =15.81), i.e. how much attention to mathematics is given by friends and family, the *teacher student relations* (mean rank C5 = 51.44 vs. C10 = 14.90), the *mathematics teacher's support* (mean rank C5 = 52.22 vs. C10 = 14.43), and the *teacher behavior: student orientation* (mean rank C5 = 54.33 vs. C10 = 15.14), respectively. No highly statistical differences were found for any other group combination.

Hence, the statistical test on a global level suggests that overall; the Europe cluster and the developing countries cluster are the most dissimilar to each other. Students in the Europe cluster have a higher economic, social and cultural status—but the students in the developing countries cluster have higher interests, more motivation to learn, and higher subjective norms from their friends and family in mathematics. Furthermore, students in the developing countries tend to report better relations with their teachers.

When comparing Finland to other countries, the rather negative attitudes towards mathematics were already observed in the 2003 assessment cycle. In both the *interest in* and the *enjoyment of* mathematics, Finland was ranked 37th out of the 40 participating countries (Linnakylä et al. 2011).

Moreover, in a longitudinal study of Finnish grade 1 to grade 12 students by Metsämuuronen et al. (2012), it was concluded that student contentment in regard to school in Finland decreases significantly from the second to the eighth grade, while it then very slightly increases again starting from the ninth grade. The majority (82%[11]) of the Finnish students participating in PISA are in the ninth grade, and almost all of the rest are in the eighth grade (16%). Hence, Finnish students are at the stage in their basic education where their self-reported attitudes towards school are very poor.

Metsämuuronen et al. (2012) suggest that these generally negative attitudes of the Finnish students towards education are due to their modesty and honesty: "Part of the explanation in Finland [...] can be the appreciation of honesty and speaking frankly [...] pupils in Finland [...] are relatively humble when they describe their knowledge. This 'humbleness' may also be reflected in attitude measurements."

10.5.3 Differences between Finland and the Other Countries within the Europe Cluster

According to the clustering result, Finland is most similar to the countries in the Europe cluster. But what are the variables that separate Finland from the countries within its own cluster? Table 10.5 summarizes the highly statistically significant variables according to which Finland is different from the remaining countries, as determined by the Wilcoxon signed-rank tests.

Table 10.5 Wilcoxon signed-rank statistics (Europe - Finland clusters).

variable	1	7	10	11	16	17	18	24
Z	-3.920	3.920	3.920	3.771	3.808	3.920	3.845	3.920
P	★★★	★★★	★★★	★★★	★★★	★★★	★★★	★★★

As can be seen from Table 10.5, the majority of the Europe cluster has a significantly lower *ESCS* than Finland ($z = -3.92$, $p < 0.001$). Nevertheless, the Europe cluster majority has a significantly higher *self-responsibility for failing in mathematics* ($z = 3.92$, $p < 0.001$), *anxiety towards mathematics* ($z = 3.771$, $p < 0.001$), and *self-efficacy in mathematics* ($z = 3.92$, $p < 0.001$) than Finland. Furthermore, the Europe cluster in general shows higher scores in many variables that measure emphasis of formal assessment and how much time students spend with studying.

[11] Own calculation on PISA 2012 data.

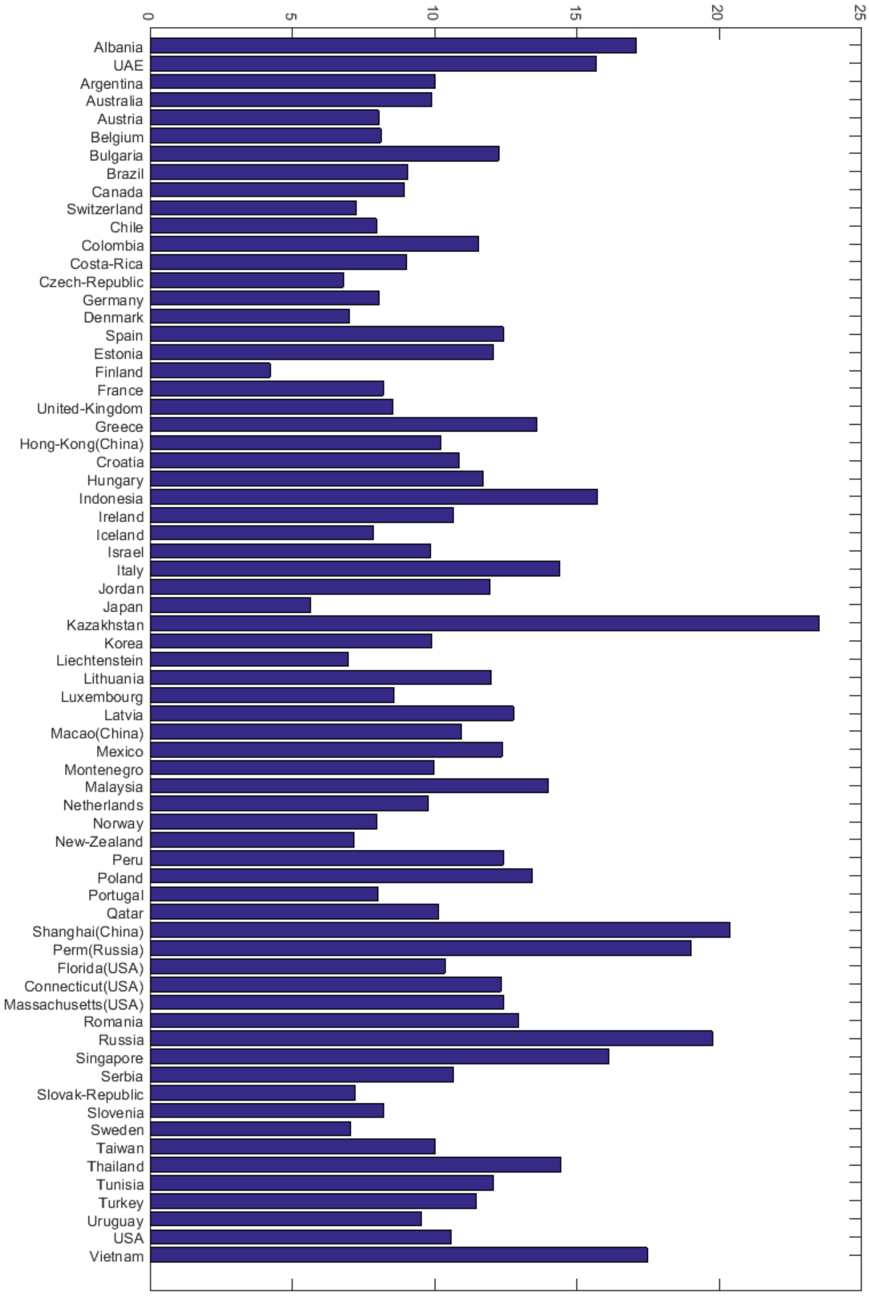


Fig. 10.7 Weighted average of out-of-school study hours for all in PISA participating countries. In comparison to all the other countries, Finnish students study the least after school.

In particular, there is a significantly higher *work ethic in mathematics* (z = 3.808, $p < 0.001$) and more *out-of-school study hours* in the Europe cluster than in Finland (z = 3.920, $p < 0.001$). The latter is illustrated in Fig. 10.7, where the weighted average out-of-school study hours for students in all participating PISA countries are plotted. As can be seen from the figure, Finnish students not only study the least outside of school within their own Europe cluster but also compared to all other countries participating in PISA.

In addition, the *learning time (min. per week) - test language* in Europe is significantly greater than in Finland (z = 3.845, $p < 0.001$, see Table 10.5), and Europe has a significantly higher score in *teacher behaviour: formative assessment* than Finland (z = 3.920, $p < 0.001$). In summary, these results support observations by Sahlberg (2011) who writes that educational decision makers in Finland "do not seem to believe that doing more of the same in education would necessarily make any significant difference for improvement."

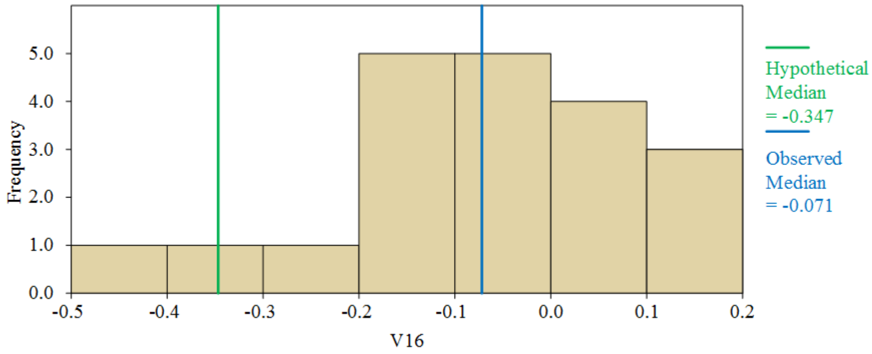


Fig. 10.8 One-Sample Wilcoxon Rank Test for work ethic: The work ethic of students in Finland is significantly lower than the work ethic of students in the European cluster.

As can be seen from the Wilcoxon signed-rank test result and Fig 10.8, 15-year-old students in Finland seem to already have a rather relaxed attitude towards formal assessment and investing time in studies. This is particularly evident in the highly statistically significantly lower work ethic[12] of Finnish students.

One should also keep in mind that the systematic apprenticeship of reading, writing, and mathematics begins later in Finland than in Europe ($z = -3.435$, $p < 0.001$). This is illustrated in Fig. 10.9. In Finland, children are seven years old when they start school. Combined with the finding that the hours of formal instruction of certain subjects are, as described in the above paragraph, significantly

[12] The *work ethics* scale index is computed with the Rasch model and by using the extent to which students agree or disagree with the following statements: *I finish my homework in time for mathematics class; I work hard on my mathematics homework; I am prepared for my mathematics exams; I study hard for mathematics quizzes; I keep studying until I understand mathematics material; I pay attention in mathematics class; I listen in mathematics class; I avoid distractions when I am studying mathematics, I keep my mathematics work well organised.*

lower in Finland, this means that Finnish students spend less time at school than students in other countries. This finding has also been emphasized by Kumpulainen & Lankinen (2012).

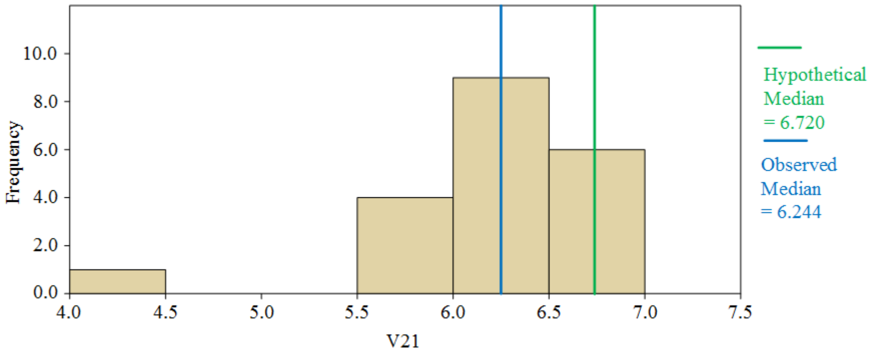


Fig. 10.9 One-Sample Wilcoxon Rank Test for *age at <ISCED 1>*: Systematic apprenticeship of reading, writing and mathematics begins significantly later in Finland than in Europe.

10.5.4 Europe Cluster in Comparison to the Nordic/English-Speaking Cluster

A Mann-Whitney U test was run to determine if there were differences in the 27 variables between the Europe and the Nordic/English-speaking clusters. Distributions of the 27 variables for the two groups were not similar, as assessed by visual inspection. The test statistics can be found in Table 10.6.

Table 10.6 Mann-Whitney U test results comparing the Europe cluster to the Nordic/English speaking cluster.

PISA variable ID	4	8	9	12	15	16	18	22	23	25
U	19	27	5	30	1	38	22	20	28	20
Z	-4.004	-3.705	-4.528	-3.593	-4.678	-3.293	-3.892	-3.967	-3.668	-3.967
p	★★★	★★★	★★★	★★★	★★★	★★★	★★★	★★★	★★★	★★★

When we combine the test results of the Mann-Whitney U test of the Nordic/English-speaking versus Europe and the Wilcoxon signed-rank test of Europe versus land, we find that two variables (16 and 18) augment Finland's special characteristics: *work ethic* and *study time (test language)* are statistically significantly lower in Europe and even lower in Finland. As described above, these variables measure how much time students spend studying and how much they strive for high grades in mathematics.

According to the Mann-Whitney U test, there was a significant ($p < 0.001$) difference in the *attitude towards school: learning activities*, the *interest in and enjoyment of mathematics*, the *instrumental motivation to learn mathematics*, the *self-concept in mathematics*, the *subjective norms in mathematics*, the *mathematics work ethic*, the *test language learning time*, the *teacher student relations*, the *mathematics teacher's support*, and the *teache*r behaviour: student orientation between the two clusters. In all of these variables, the Nordic/English-speaking cluster showed higher values than the Europe cluster. With reference to Table 10.6, the *subjective norms in mathematics* seems to be the most important variable that separates the Nordic/English-speaking from the Europe cluster.

The comparisons of the Nordic/English cluster to the European cluster mostly revealed variables that estimate the students' own perception of their merits and importance. It is especially interesting that the self-reported self-concept is significantly lower in Finland because this PISA 2012 variable actually explains the performance of Finnish students in the PISA mathematics test fairly well, and it is the mathematics scale index that correlates the most with their plausible values in mathematics (Saarela & Kärkkäinen 2014). However, it seems that even if the Finnish students evaluate their own skills realistically, they are more modest about them. Generally, students in the Nordic/English cluster tend to have higher opinions about themselves, are more motivated, and report better relations to their teachers.

The average mathematics performance based on the plausible values of the countries in the Nordic/English-speaking cluster is 495.3, while the mean mathematics performance of the countries in the European cluster is higher (500.5). We conclude that learning time and positive student-teacher relations seem to be less important features than collaborative skills or being free from ostentation for explaining students' success in the PISA test.

10.6 Visual LA of the PISA Results

The macro-level LA of Finnish basic educational system, through the lens of background, PISA and our empirical analysis, is visualized in the dashboard of Fig. 10.10. This dashboard consists of four panels and its composition was inspired by Ferguson & Shum (2012).

Finland has been a top performing PISA country in the last five assessment cycles (top-left panel), although the ranking especially in mathematics clearly decreased in 2012. Interesting success factors of the educational system are the cultural deviations from world's midlevel as feminine culture and with a low power distance, according to the Hofstede model (top-right panel). The system is based on the strong autonomy and authority of the highly educated teachers, with small amount of formal assessment and, especially, complete lack of national comparative assessments of the learning results (bottom-left panel). Also rich common curriculum for untracked groups of students, who start late their systematic apprenticeship in reading, mathematics, and science, is present. As a whole, equity and equality characterize the system, which provides strong student support, e.g., in the form of free lunches, health care, school transportation (bottom-left).

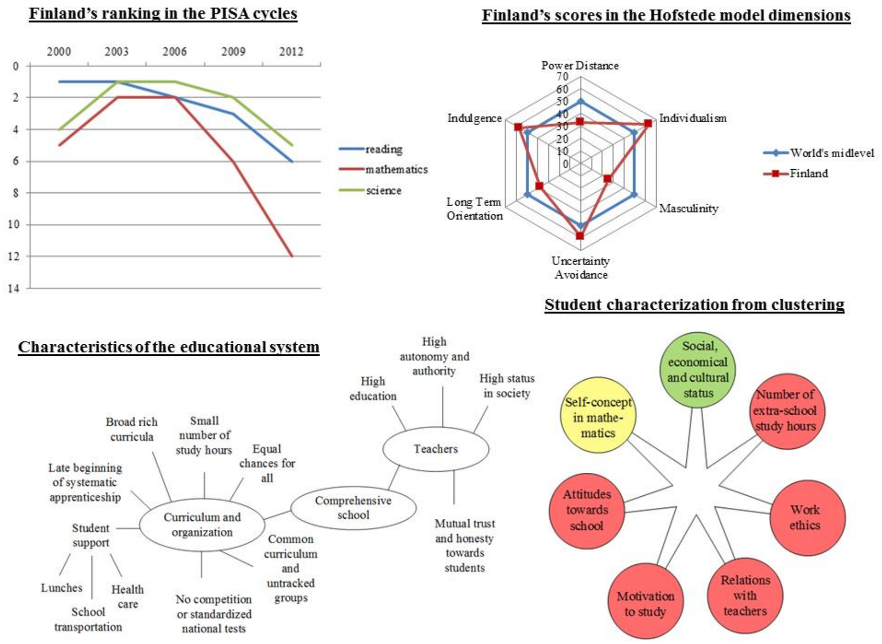


Fig. 10.10. PISA dashboard for Finland (inspired by Ferguson & Shum 2012)

However, many contradicting factors about the Finnish students in relation to their high PISA results emerged in the empirical LA analysis (bottom-right panel): they have low motivation to learn and excel in school, low interest in school topics, low work ethics, and exceptionally small number of extra-school study hours. Importance of studies and, especially, mathematics are considered low for the future career. The overall evaluation of the different facets of the dashboard indicates that the lowering trend of PISA and especially mathematics performance of the Finnish students may continue. To improve the system, perhaps again as number one ranked in PISA, students need to be more motivated and oriented towards schoolwork, extra-school study hours, and mathematics with future career orientation clearly in mind. We also hypothesize that the complete common, joint, and untracked subject orientations demotivate the most talented students by requiring minimal efforts from them. All this, then, provides further challenges to the upper secondary and higher education afterwards.

10.7 Discussion

Let us next briefly summarize the empirical findings from the previous sections. These were obtained by utilizing one of the illuminated educational clustering techniques, hierarchical clustering, and by taking into account all the specific demands of PISA data discussed above. As suggested by the Davies-Bouldin cluster validation index, we first divided the students of all the PISA-participating countries into ten separate groups. These found clusters could be explained by the culture and geographical location of the countries in them. Nevertheless, Finland surprisingly belonged to the Europe cluster (see Fig. 10.10), while all the other Scandinavian countries belonged to the cluster of Nordic/English-speaking countries. Hence, this illustrates how similar educational systems (see Bulle 2011) can be reflected by different student characterizations.

Statistical significance tests of the clustering result revealed why particular countries were allocated to a certain cluster. At first, it seemed that the results of the statistical test were somehow contradictory as better performing countries had worse student teacher relations—and generally showed less confidence in their own achievements and skills. Moreover, the work ethic of the students in the better performing Europe cluster was significantly lower than that of the students in the Nordic/English-speaking countries cluster—and the better performing Finnish students showed an even significantly worse work ethic than the remaining students in the Europe cluster. However, these findings seem to be connected and explicable by the existing research related to the Finnish culture in general.

As was explained in the literature review about the Finnish educational system and culture, Finnish citizens are modest about their own achievements, and they place great emphasis on equity and equality. The most important driving factors in the life of this, according to Hofstede's (2011) model, highly feminine country are to live a good life and to care for others rather than focusing on one's own success and desire to be the best. This is interesting because, as emphasized in our literature review, French et al. (2015) found a negative causal relationship between education expenditures and power distance and masculinity: The less masculine the country, the higher was the education expenditure. Furthermore, Finnish students seem to have an extremely relaxed attitude towards formal assessment and investing time in studies, as can be expected in a feminine country.

Finally, the main success of Finnish students in PISA seems to a great extent to be related to the— in comparison with other countries—relatively better scores of the lowest scoring Finnish students (Andersen 2010), which in turn is supported again by the collaborative and ostentation-free thinking in the country. However, as illustrated in the top-left panel of Fig. 10.10, Finland's ranking significantly dropped in the latest PISA 2012 assessment (OECD 2013*b*), and according to the overall characterization of the Finnish students as just given and visualized in the bottom-right panel of Fig. 10.10, the negative trend in performance might have continued in PISA 2015[13].

[13] Data from the PISA 2015 will be published by the OECD in December 2016 (National Center for Education Statistics 2016).

10.8 Conclusions

LA is a growing and expanding research field. Traditionally, many studies have concentrated on analyzing educational data originating from a macro- or at the most, meso-level. The publicly available and high quality PISA data sets, on the other hand, provide the opportunity to conduct big data LA research on the macro-level, because they consist of data of a whole population of international students.

In this chapter, we have introduced the background for conducting large-scale LA research on PISA. We have described the main data sets—as well as the complexities within them—and have discussed how to work with these data. Moreover, we have provided a review of relevant clustering studies within the educational domain. Our empirical work, as discussed in the previous section, provided novel findings and strengthened earlier knowledge on the particularities of the Finnish educational system that has obtained much attention during the 21st century due to the exceptionally good performance of the Finnish students in the PISA tests.

We used quantitative LA methods to identify the main attributes of individual learners affecting their learning experience in the environment where the learning occurs (Fournier et al. 2011). Similarly to the reviewed educational clustering studies in Section 10.2.2, we were analyzing the student model, but differently from these, our model represented a prototype of a national student population obtained by weighted aggregation. Concerning Finland, the high-achieving country inside PISA assessments, it was concluded that an educational system promoting student collaboration, unassumingness, and equity can successfully cope with the challenges of negative attitudes towards mathematics, low work ethic, and little study time outside school. This summarizes the evidence-based knowledge discovered about the long-term impact of educational policies and practices on the achievement targets (Piety et al. 2014). Such a conclusion provides also an example of national education system assessment using big data LA as illustrated in Fig. 10.1: The international objectives driven data collection and transformation improves understanding of educational arrangements via proper analysis methods that are able to cope with the specialties of the sampled large-scale data.

The big data LA as described in Section 10.1 and depicted in Fig. 10.1, linking together the four dimensions of LA proposed by Chatti et al. (2014) (see also Greller & Drachsler 2012), encapsulated and supported the overall management of the large-scale educational system assessment based on the PISA data. Our empirical work exemplifies the multiple facets of LA: hierarchical clustering as a data mining technique, visualization of the dendrogram to illustrate the clustering result, and statistical testing to verify the findings. Thus, our work increased the awareness on the macro-level of educational systems. We promoted reflection of the main characteristics that differentiate the students in various educational environments, according to the objectives of LA by Chatti et al. (2014) (see Section 3.3). Our reflections of the PISA results were emphasized in the dashboard Fig. 10.10 using different LA visualization tools. This dashboard facilitates awareness and monitoring of critical educational aspects for the Finnish 15-year-old student population (Beheshitha et al. 2016).

As a whole, PISA—as well as the other large-scale-assessments, such as those mentioned in Section 10.3.4—provide a very rich and interesting source for macro-level LA studies. We think that the methods and the framework developed for the publicly available large-scale assessment data sets can and will advance the open architecture of educational applications, which Peña-Ayala (2014) has identified as one of the shortcomings of the current educational data analysis research area.

As part of the future research, we intend to repeat our study using the individual students instead of the country-level aggregation as data for clustering. Furthermore, one of the recent trends in LA focuses on educational process mining (Sedrakyan et al. 2016, Mukala et al. 2015, Trčka et al. 2010). For the traditional pen-and-paper PISA tests, this is not an option. However, for the future PISA cycles, where the tests will be increasingly conducted electronically, and where log event data will therefore be available (compare the PISA 2012 problem-solving test, which was conducted electronically, and where log files can be downloaded from the above-cited OECD webpage), this would provide an interesting and promising direction for future research.

Acknowledgements

The authors would like to thank Dr. Jouni Välijärvi, Dr. Kari Nissinen, and Dr. Antero Malin from the Finnish Institute for Educational Research for the valuable discussions and support concerning the PISA data and assessment. We also acknowledge BSc Susanne Jauhiainen for the technical assistance in preparing the final version of the chapter.

References

Allen LK, Mills C, Jacovina ME, Crossley S, D'Mello S, McNamara DS (2016) Investigating Boredom and Engagement During Writing Using Multiple Sources of Information: The Essay, the Writer, and Keystrokes. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 114–123

Almeda MV, Scupelli P, Baker RS, Weber M, Fisher A (2014) Clustering of design decisions in classroom visual displays. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 44–48

Andere E (2015) Are Teachers Crucial for Academic Achievement? Finland Educational Success in a Comparative Perspective. Education Policy Analysis Archives 23(39):1–27

Andersen FØ (2010) Danish and Finnish PISA results in a comparative, qualitative perspective: How can the stable and distinct differences between the Danish and Finnish PISA results be explained? Educational Assessment, Evaluation and Accountability 22(2):159–175

Auld E, Morris P (2016) PISA, policy and persuasion: translating complex conditions into education 'best practice'. Comparative Education 52(2):202–229

Bank V (2012) On OECD policies and the pitfalls in economy-driven education: The case of Germany. Journal of Curriculum Studies 44(2):193–210

Beheshitha SS, Hatala M, Gašević D, Joksimović S (2016) The Role of Achievement Goal Orientations When Studying Effect of Learning Analytics Visualizations. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 54–63

Bergner Y, Colvin K, Pritchard DE (2015) Estimation of Ability from Homework Items when There Are Missing and/or Multiple Attempts. In: Proceedings of the Fifth International Conference on Learning Analytics & Knowledge, ACM, pp 118–125

Blikstein P, Worsley M, Piech C, Sahami M, Cooper S, Koller D (2014) Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. Journal of the Learning Sciences 23(4):561–599

Bogarin A, Romero C, Cerezo R, Sanchez-Santillan M (2014) Clustering for Improving Educational Process Mining. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 11–15

Bouchet F, Harley JM, Trevors GJ, Azevedo R (2013) Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning. Journal of Educational Data Mining 5(1):104–146

Brown MG, DeMonbrun RM, Lonn S, Aguilar SJ, Teasley SD (2016) What and when: The Role of Course Type and Timing in Students' Academic Performance. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 459–468

Bulle N (2011) Comparing OECD educational models through the prism of PISA. Comparative Education 47(4):503–521

Chandra E, Nandhini K (2010) Knowledge mining from student data. European Journal of Scientific Research 47(1):156–163

Chatti MA, Dyckhoff AL, Schroeder U, Thüs H (2012) A Reference Model for Learning Analytics. International Journal of Technology Enhanced Learning 4(5–6):318–331

Chatti MA, Lukarov V, Thues H, Muslim A, Yousef AMF, Wahid U, Greven C, Chakrabarti A, Schroeder U (2014) Learning Analytics: Challenges and Future Research Directions. E-learning and Education (Eleed) Journal 10:1–16

Coffrin C, Corrin L, de Barba P, Kennedy G (2014) Visualizing patterns of student engagement and performance in MOOCs. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, ACM, pp 83–92

Colthorpe K, Zimbardi K, Ainscough L, Anderson S (2015) Know Thy Student! Combining Learning Analytics and Critical Reflections to Increase Understanding of Students Self-Regulated Learning in an Authentic Setting. Journal of Learning Analytics 2(1):134–155

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1(2):224–227

Dawson S, Gašević D, Siemens G, Joksimović S (2014) Current state and future trends: A citation network analysis of the learning analytics field. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 231–240

Dawson S, Gašević D, Mirriahi N (2015) Challenging assumptions in learning analytics. Journal of Learning Analytics 2(3):1–3

Deng Z, Gopinathan S (2016) PISA and high-performing education systems: explaining Singapores education success. Comparative Education 52(4):449–472

Desmarais MC, Lemieux F (2013) Clustering and Visualizing Study State Sequences. In: Proceedings of the 6th International Conference on Educational Data Mining, pp 224–227

Drabowicz T (2014) Gender and digital usage inequality among adolescents: A comparative study of 39 countries. Computers & Education 74:98 – 111

Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6(3):241– 252

Embretson SE, Reise SP (2013) Item Response Theory. Psychology Press

Erdogdu F, Erdogdu E (2015) The impact of access to ICT, student background and school/home environment on academic success of students in Turkey: An international comparative analysis. Computers & Education 82:26 – 49

Ferguson R (2012) Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5–6):304–317

Ferguson R, Shum SB (2012) Social Learning Analytics: Five Approaches. In: Proceedings of the Second International Conference on Learning Analytics & Knowledge, ACM, pp 23–33

Ferguson R, Macfadyen L, Clow D, Tynan B, Alexander S, Dawson S (2014) Setting Learning Analytics in Context: Overcoming the Barriers to Large-Scale Adoption. Journal of Learning Analytics 1(3):120–144

Ferguson R, Clow D (2015) Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In: Proceedings of the Fifth International Conference on Learning Analytics & Knowledge, ACM, pp 51–58

Fonseca J, Valente MO, Conboy J (2011) Student characteristics and PISA science performance: Portugal in cross-national comparison. Procedia-Social and Behavioral Sciences 12:322–329

Fournier H, Kop R, Sitlia H (2011) The Value of Learning Analytics to Networked Learning on a Personal Learning Environment. In: Proceedings of the First International Conference on Learning Analytics & Knowledge, pp 104–109

French JJ, French A, Li WX (2015) The relationship among cultural dimensions, education expenditure, and PISA performance. International Journal of Educational Development 42:25–34

Grawemeyer B, Mavrikis M, Holmes W, Gutierrez-Santos S, Wiedmann M, Rummel N (2016) Affecting Off-task Behaviour: How Affect-aware Feedback Can Improve Student Learning. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 104–113

Gray G, McGuinness C, Owende P, Carthy A, (2014) A Review of Psychometric Data Analysis and Applications in Modelling of Academic Achievement in Tertiary Education. Journal of Learning Analytics 1(1):75–106

Grek S (2009) Governing by numbers: The PISA effect in Europe. Journal of Education Policy 24(1):23–37

Greller W, Drachsler H (2012) Translating Learning into Numbers: A Generic Framework for Learning Analytics. Educational Technology & Society 15(3):42–57

Gupta D, Sharma A, Unny N, Manjunath G (2014) Graphical analysis and visualization of big data in business domains. In: Big Data Analytics, Lecture Notes in Computer Science (8883), Springer-Verlag, pp 53–56

Hansen JD, Reich J (2015) Socioeconomic status and MOOC enrollment: enriching demographic information with external datasets. In: Proceedings of the Fifth International Conference on Learning Analytics & Knowledge, ACM, pp 59–63

Hecking T, Chounta IA, Hoppe HU (2016) Investigating social and semantic user roles in MOOC discussion forums. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 198–207

Hershkovitz A, Knight S, Dawson S, Jovanović J, Gašević D (2016) About" Learning" and" Analytics. Journal of Learning Analytics 3(2):1–5

Hickey DT, Kelley TA, Shen X (2014) Small to Big Before Massive: Scaling up Participatory Learning Analytics. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 93–97

Hitlin S, Piliavin J (2004) Values: Reviving a dormant concept. Annual Review of Sociology 30:359–393

Hofstede G (2011) Dimensionalizing cultures: The Hofstede model in context. Online readings in psychology and culture 2(1):8

Hu X, Zhang Y, Chu SKW, Ke X (2016) Towards Personalizing an e-Quiz Bank for Primary School Students: An Exploration with Association Rule Mining and Clustering. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 25–29

Joksimović S, Manataki A, Gašević D, Dawson S, Kovanović V, de Kereki IF (2016) Translating Network Position into Performance: Importance of Centrality in Different Network Configurations. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 314–323

Kärkkäinen T, Saarela M (2015) Robust principal component analysis of data with missing values. In: Lecture Notes in Artificial Intelligence (9166), Springer International Publishing, pp 140–154

Kriegbaum K, Jansen M, Spinath B (2015) Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. Learning and Individual Differences 43:140–148

Kruskal W, Wallis W (1952) Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 47(260):583–621

Kumpulainen K, Lankinen T (2012) Striving for educational equity and excellence. In: Miracle of education, Springer, pp 69–81

Laney D (2001) 3D data management: Controlling data volume, velocity and variety. Tech. rep., META Group

Li N, Cohen WW, Koedinger KR (2013) Discovering student models with a clustering algorithm using problem content. In: Proceedings of the 6th International Conference on Educational Data Mining, pp 98–105

Linnakylä P, Välijärvi J, Arffman I (2011) Finnish Basic Education – When Equity and Excellence Meet. In: Equity and Excellence in Education: Towards maximal learning opportunities for all students, Routledge, New York, pp 190–214

Long P, Siemens G (2011) Penetrating the fog: Analytics in learning and education. EDUCAUSE Review 46(5):30–40

López MI, Luna JM, Romero C, Ventura S (2012) Classification via clustering for predicting final marks based on student participation in forums. In: Proceedings of the 5th International Conference on Educational Data Mining, pp 148–151

Marsman M (2014) Plausible Values in Statistical Inference. Universiteit Twente

Merceron A, Blikstein P, Siemens G (2016) Learning Analytics: From Big Data to Meaningful Data. Journal of Learning Analytics 2(3):4–8

Metsämuuronen J, Svedlin R, Ilic J (2012) Change in pupils' and students' attitudes toward school as a function of age - A Finnish perspective. Journal of Educational and Developmental Psychology 2(2):134–151

Monseur C, Adams R (2008) Plausible Values: How to Deal with Their Limitations. Journal of Applied Measurement 10(3):320–334

Morgan H (2014) Review of Research: The Education System in Finland: A Success Story Other Countries Can Emulate. Childhood Education 90(6):453–457

Mukala P, Buijs J, Leemans M, van der Aalst W (2015) Learning analytics on Coursera event data: A process mining approach

Musik A (2016) Philologenverband bezeichnet Pisa-Studie als Geldverschwendung. http://www.deutschlandfunk.de/bildungsforschung-in-der-kritikphilologenverband.680.de.html?dram:articleid = 347675

National Center for Education Statistics (2016) Program for International Student Assessment. https://nces.ed.gov/surveys/pisa/

OECD (2011) Finland: Slow and Steady Reform for Consistently High Results. In: Successful Reformers in Education: Lessons from PISA for the United States, OECD, pp 117–135

OECD (2012) PISA 2009 Technical Report. OECD Publishing

OECD (2013a) PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III). PISA, OECD Publishing

OECD (2013b) PISA 2012 Results: What Students Know and Can Do (Volume I) Student Performance in Mathematics, Reading and Science: Student Performance in Mathematics, Reading and Science. v. 1, OECD Publishing

OECD (2014a) PISA 2012 Results in Focus: What 15-year-olds know and what they can do with what they know. OECD Publishing. Paris: France

OECD (2014b) PISA 2012 Technical Report

Olsen RV (2005a) Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science. PhD thesis, University of Oslo

Olsen RV (2005b) An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension? Nordic Studies in Science Education 1(1):81–94

Pardo A, Teasley S (2014) Learning Analytics Research, Theory and Practice: Widening the Discipline. Journal of Learning Analytics 1(3):4–6

Peña-Ayala A (2014) Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications 41(4):1432 – 1462

Picciano AG (2012) The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks 16(3):9–20

Piety PJ, Hickey DT, Bishop M (2014) Educational Data Sciences - Framing Emergent Practices for Analytics of Learning, Organizations, and Systems. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 193–202

Rasch G (1960) Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Rasmussen J, Bayer M (2014) Comparative study of teaching content in teacher education programmes in Canada, Denmark, Finland and Singapore. Journal of Curriculum Studies 46(6):798–818

Rogers T (2015) Critical Realism and Learning Analytics Research: Epistemological Implications of an Ontological Foundation. In: Proceedings of the Fifth International Conference on Learning Analytics & Knowledge, ACM, pp 223–230

Reich J, Tingley DH, Leder-Luis J, Roberts ME, Stewart B (2014) Computer-assisted reading and discovery for student generated text in massive open online courses. Journal of Learning Analytics 2(1):156–184

Rutkowski L, Gonzalez E, Joncas M, von Davier M (2010) International Large-Scale Assessment Data Issues in Secondary Analysis and Reporting. Educational Researcher 39(2):142–151

Saarela M, Kärkkäinen T (2014) Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In: Proceedings of the 7th International Conference on Educational Data Mining, pp 60–68

Saarela M, Kärkkäinen T (2015a) Analysing Student Performance using Sparse Data of Core Bachelor Courses. Journal of Educational Data Mining 7(1):3–32

Saarela M, Kärkkäinen T (2015b) Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In: Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015), pp 156–163

Saarela M, Kärkkäinen T (2015c) Weighted Clustering of Sparse Educational Data. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp 337-342

Saarela M, Kärkkäinen T, Lahtonen T, Rossi T (2016a) Expert-based versus citation-based ranking of scholarly and scientific publication channels. Journal of Informetrics 10(3):693 – 718

Saarela M, Yener B, Zaki MJ, Kärkkäinen T (2016b) Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach. In: MLDEAS workshop of the 33rd International Conference on Machine Learning, pp 1-8

Sahlberg P (2011) Finnish lessons. Teachers College Press

Santos JL, Klerkx J, Duval E, Gago D, Rodriıguez L (2014) Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 98–102

Schatz M, Popovic A, Dervin F (2016) From PISA to national branding: exploring Finnish education. Discourse: Studies in the Cultural Politics of Education pp 1–13

Sedrakyan G, Weerdt JD, Snoeck M (2016) Process-mining enabled feedback: Tell me what I did wrong vs. tell me how to do it right. Computers in Human Behavior 57:352 – 376

Segedy JR, Kinnebrew JS, Biswas G (2015) Using Coherence Analysis to Characterize Self-Regulated Learning Behaviours in Open-Ended Learning Environments. Journal of Learning Analytics 2(1):13–48

Siemens G (2013) Learning analytics: The emergence of a discipline. American Behavioral Scientist 57:1380–1400

Siemens G (2014) The Journal of Learning Analytics: Supporting and Promoting Learning Analytics Research. Journal of Learning Analytics 1(1):3–5

Siemens G, Baker RS (2012) Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: Proceedings of the 2nd International Conference on Learning Analytics & Knowledge, ACM, pp 252–254

Simola H (2005) The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. Comparative Education 41(4):455–470

Skryabin M, Zhang J, Liu L, Zhang D (2015) How the ICT development level and usage influence student achievement in reading, mathematics, and science. Computers & Education 85:49–58

Tømte C, Hatlevik O (2011) Gender-differences in Self-efficacy ICT related to various ICT-user profiles in Finland and Norway. How do self-efficacy, gender and ICT-user profiles relate to findings from PISA 2006. Computers & Education 57(1):1416 – 1424

Trčka N, Pechenizkiy M, van der Aalst W (2010) Process mining from educational data. Chapman & Hall/CRC

Välijärvi J, Kupari P, Linnakylä P, Reinikainen P, Sulkunen S, Törnroos J, Arffman I (2007) The Finnish success in PISA - and some reasons behind it: PISA 2003. Jyväskylän yliopisto, Koulutuksen tutkimuslaitos

Verbert K, Manouselis N, Drachsler H, Duval E (2012) Dataset-Driven Research to Support Learning and Knowledge Analytics. Educational Technology & Society 15(3):133–148

Vogelsang T, Ruppertz L (2015) On the validity of peer grading and a cloud teaching assistant system. In: Proceedings of the Fifth International Conference on Learning Analytics & Knowledge, ACM, pp 41–50

Waldow F, Takayama K, Sung YK (2014) Rethinking the pattern of external policy referencing: media discourses over the Asian Tigers: PISA success in Australia, Germany and South Korea. Comparative Education 50(3):302–321

Wang Y, Paquette L, Baker R (2014) A longitudinal study on learner career advancement in MOOCs. Journal of Learning Analytics 1(3):203–206

Wise AF, Shaffer DW (2015) Why Theory Matters More than Ever in the Age of Big Data. Journal of Learning Analytics 2(2):5–13

Wise AF, Cui Y, Vytasek J (2016) Bringing order to chaos in MOOC discussion forums with content-related thread identification. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, ACM, pp 188–197

Worsley M, Blikstein P (2014) Analyzing Engineering Design through the Lens of Computation. Journal of Learning Analytics 1(2):151–186

Wu M (2005) The role of plausible values in large-scale surveys. Studies in Educational Evaluation 31(2):114–128

Wu M, Adams R (2002) Plausible values: Why they are important. In: 11th International Objective Measurement Workshop, New Orleans

Xing W, Wadholm B, Goggins S (2014) Learning analytics in CSCL with a focus on assessment: An exploratory study of activity theory-informed cluster analysis. In: Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM, pp 59–67

Yates L (2013) Revisiting curriculum, the numbers game and the inequality problem. Journal of Curriculum Studies 45(1):39–51

Ye C, Biswas G (2014) Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information. Journal of Learning Analytics 1(3):169–172

Zaki MJ, Meira Jr W (2014) Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press

Zhong ZJ (2011) From access to usage: The divide of self-reported digital skills among adolescents. Computers & Education 56(3):736 – 746

PIII

DISCOVERING GENDER-SPECIFIC KNOWLEDGE FROM FINNISH BASIC EDUCATION USING PISA SCALE INDICES

by

Mirka Saarela, Tommi Kärkkäinen 2014

Proc. of the 7th International Conference on Educational Data Mining, pp. 60–67

Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices

Mirka Saarela
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
mirka.saarela@gmail.com

Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
tommi.karkkainen@jyu.fi

ABSTRACT

The Programme for International Student Assessment, PISA, is a worldwide study to assess knowledge and skills of 15-year-old students. Results of the latest PISA survey conducted in 2012 were published in December 2013. According to the results, Finland is one of the few countries where girls performed better in mathematics than boys. The purpose of this work is to refine the analysis of this observation by using education data mining techniques. More precisely, as part of standard PISA preprocessing phase certain scale indices are constructed based on information gathered from the background questionnaire of each participating student. The indices describe, e.g., students' engagement, drive and self-beliefs, especially related to mathematics, the main assessment area in PISA 2012. However, around 30% of the scale indices are missing so that a nonstructured sparsity pattern must be dealt with. We handle this using a special, robust clustering technique, which is then applied to Finnish subset of PISA data. Already direct interpretation of the created clusters reveals interesting patterns. Clusterwise analysis through relationship mining refines the confidence on our final conclusion that attitudes towards mathematics which are often gender-specific are the most important factors to explain the performance in mathematics.

Keywords

PISA, robust clustering, frequent itemset, association rule

1. INTRODUCTION

PISA (Programme for International Student Assessment) is an international assessment programme by the Organisation for Economic Co-operation and Development (OECD) that studies students' learning outcomes in reading, mathematics, and scientific literacy triennially. It is referred as the "world's premier yardstick for evaluating the quality, equity and efficiency of school systems" [21]. More than seventy countries and economies have already participated in PISA. Finland has consistently been one of the top-performing countries in the assessment [11]. Each time the study is repeated the main learning outcome focus area changes. In the latest assessment (PISA 2012) it was mathematics. A database of the results is publicly available[1].

One general key finding from PISA 2012 was the gender difference in mathematics performance: On average, boys outperform girls in mathematics. Finland, however, is, according to the assessment, one of the eight countries where girls perform better than boys in mathematics: The mean score of girls in mathematics was 520 while boys had the mean score of 517 [23]. Despite the slightly better performance in mathematics women are, also in Finland, underrepresented in mathematics related jobs [28].

The purpose of this work is to apply educational data mining approch and corresponding techniques to study the performance of Finnish student population in mathematics, focusing especially on gender-related findings. As part of standard PISA preprocessing phase, certain *scale indices* are constructed based on information gathered from the background questionnaire for each participating student [21]. These indices describe, e.g., students' engagement, drive and self-beliefs, especially related to mathematics. However, around 30% of the scale indices are missing due to lack of reliable student responses for the background questions. This means that the knowledge discovery process is realized with data having a nonstructured sparsity pattern. We handle this using a special, robust clustering technique as proposed in [4]. Furthermore, the clustering result obtained is further analyzed using itemset mining [1] to foster the generation of novel information and new knowledge.

The contents of the paper is as follows: First, we provide a short summary on PISA data and how students' capabilities and attributes are presented. We then describe a certain set of scale index variables that are associated with the performance in mathematics. Subsequently, we apply methods from two (see [7] for a complete categorization) of the main branches in educational data mining. In Section 3, we utilize a special clustering approach to find groups of students with similar characteristics with respect to scale indices. In order to further refine the characterization of student groups, we then apply frequent itemset mining and association rule learning to selected clusters in Section 4. Finally, we sum-

[1] See `http://www.oecd.org/pisa/pisaproducts/`.

marize and conclude our study in Section 5.

2. ON PISA DATA

We apply educational data mining for the PISA 2012 data subset of Finland. In each country participating PISA, the schools and students selected for the survey reflect the whole population and characteristics of the educational context. In Finland, 311 schools and 10157 students from these schools were sampled for the assessment in 2012. Out of the sampled students 8829 participated in the actual PISA test. Hereby, each student that takes part has to (i) solve a set of cognitive items/tasks and (ii) fill out one background questionnaire[2] with demographic questions.

Finnish PISA data is stored in two different data sets: One data set includes all the students that participated in the test, and the second one includes all sampled schools. The student data set has more than 600 variables. A set of those variables directly encode the students answers given in the background questionnaire. Moreover, since the participating students should reflect all 15-year-old students in Finland, certain weights are assigned to each student to align the sample with the true population. In PISA reports and learning analysis, student abilities are not given as direct responses to task questions but in the form of the so-called *Plausible Values* (PVs).

Since a very broad domain of knowledge and skills should be tested but the testing time for each student is limited, only certain subsample of students respond to each item/task. In order to reliably compare results of different students, even if they have not answered exactly to the same set of items, PISA uses a generalized form of the Rasch Model [19]. Depending on how many students have solved a task correctly, a certain ”difficulty value” is assigned to each tasks and depending on how many tasks a student solved, a certain ”competence value” is assigned to each student. PVs are estimated based on difficulty and competence scores and then scaled so that the OECD average in each domain (mathematics, reading and science) is 500 and the standard deviation is 100.

Usually, five PVs are drawn from each student’s competence distribution for each main assessment area to describe the performance. For instance, in the Finnish data set for 2012 we have have five PVs for each student in reading, science, and mathematics. Moreover, since mathematics was the main assessment area, five PVs for each of the 7 subscales, i.e. subtopics in mathematics (change and relationship, quantity, space and shape, uncertainty and data, formulate, employ, interpret) are enclosed.

2.1 PISA Scale Indices

PISA scale indices (see Table 1) are derived variables based on information gathered from the background questionnaires. The scale indices are constructed in order to better characterise students dispositions, behaviours, and self-beliefs. Indeed, many of the self-reported indicators of engagement in school are strongly associated with the performance in mathematics. Especially, the *index of economic, social and cultural status* (ESCS) explains 46% of the performance variation among OECD countries so that a socio-economically more advantaged student scores 39 points higher in mathematics[3] than a less advantaged student [20]. Furthermore, according to [19], the ESCS is the ”strongest single factor associated with performance in PISA”.

Table 1 provides an overview of the PISA scale indices used in this study. In the first two columns, we provide the name of the index and it’s abbreviation used in the data set. It should be noted that some indices emphasize negative orientation with respect to mathematics. For example, it usually is not beneficial to the performance in mathematics if a student has a high value in the index which measures the anxiety towards mathematics (ANXMAT). Each index in the PISA data is standardized to have mean zero and scaled to have standard deviation one across OECD countries. Hence, a positive score index does not necessarily mean that a student has replied positively to the corresponding questions but that the answers are above the OECD average.

Correlations between the scale indices and the overall performance in mathematics are provided in the third column in Table 1. In the fourth column, ranking of the correlations based on their absolute values is given. We notice that the three indices having highest linear relationship with performance in mathematics are mathematics specific whereas the fourth index in ranking describes readiness for problem solving, and only the fifth one is the already mentioned status indicator ESCS. The correlations are computed using the subset of Finnish students for which a particular index is available. In order to obtain reliable estimates we have, as recommended in [19], analyzed each PV separately. This means that we have first computed five correlation coefficients and then used their mean as the actual result.

As already observed, not every student in the data set has a value for each of the indices. In fact, 33.24% of the index values are missing/invalid. There are different reasons why a specific scale index for a particular student is unusable. First of all, not all background questions were administered to all students. Students, that were not administered the questions included in the index had missing value by design. Second of all, it might be that the student got the questions but did not answer them. Finally, a reason for a missing index value can be that questions were answered but answers were found to be unreliable or invalid in manual scanning.

3. CLUSTER ANALYSIS USING ROBUST PROTOTYPES

Clustering is an unsupervised data analysis technique, where a given set of objects is divided into subsets (clusters) such that objects in the same cluster are similar to each other and dissimilar to objects in other clusters. Even if this appears as a simple rule, there are many approaches for clustering [10]. The classical division of algorithms is the separation into *partitional* and *hierarchical* clustering methods [16, 29]. Hierarchical clustering is usually applied for small data sets since most of the algorithms have quadratic or higher computational complexity [9]. However, the main difference be-

[2] An example of such background questionnaire can be found from `http://nces.ed.gov/surveys/pisa/pdf/MS12_StQ_FormA_ENG_USA_final.pdf`.

[3] 39 score points equal nearly one year of schooling.

Table 1: PISA scale indices and correlation to mathematics performance

PISA scale index	abbreviation	corr	rank
economic, social and cultural status	ESCS	0.36	5
sense of belonging	BELONG	0.01	15
attitude towards school: learning outcome	ATSCHL	0.15	11
attitude towards school: learning activities	ATTLNACT	0.08	12
perseverance	PERSEV	0.31	6
openness to problem solving	OPENPS	0.42	4
self-responsibility for failing in mathematics	FAILMAT	-0.20	10
interest in mathematics	INTMAT	0.25	7
instrumental motivation to learn mathematics	INSTMOT	0.23	9
self-efficacy in mathematics	MATHEFF	0.51	2
anxiety towards mathematics	ANXMAT	-0.44	3
self-concept in mathematics	SCMAT	0.52	1
behaviour in mathematics	MATBEH	0.04	13
intentions to use mathematics	MATINTFC	0.23	8
subjective norms in mathematics	SUBNORM	-0.02	14

tween these methods is related to the shape of clusters which is readily amplified in the interpretation of the clustering result. Hierarchical clustering is based on connecting locally similar objects so that the global shape of a cluster can be almost arbitrary. Partitional methods, which rely on creating subsets with respect to global similarities, are quaranteed to produce geometrically closed subsets. Moreover, the special prototype characterizing the properties of all the cluster members provides a well-defined pattern for the interpretation of the clustering result.

Prototype-based partitional clustering methods, such as *k-means*, a popular algorithm utilized also in many EDM studies [30], can be described using an iterative relocation algorithmic skeleton with an explicitly defined score function [12] (see Algorithm 1). Partitional clustering creates a $k-$partion $C = \{C_1, ..., C_k\}$ $(k \leq n)$ of data $\mathbf{X}$, such that

1) $C_i \neq \emptyset$ with $i = 1, ..., k$;
2) $\bigcup_{i=1}^{k} C_i = \mathbf{X}$; and
3) $C_i \bigcap C_j = \emptyset$ with $i, j = 1, ..., k$ and $i \neq j$.

In order to realize a prototype-based partitative clustering algorithm some further issues need to be addressed. First of all, all iterative relocation algorithms search better partitions locally so that the final result depends on the initialization. Although a lot of work has been attributed to this problem, still no universal method for identifying the initial partition exists (actually such an approach would provide an approximate solution to the clustering problem itself). Another main issue is to define the similarity measure that reflects the closedness in the data space. To this end, the amount of clusters must be determined in order to end up with one, final clustering result for the interpretation.

Our data to be clustered is problematic, because there is an arbitrary pattern of missing scale indices to deal with. Such missing values could be considered as extreme outliers because they can have any value from each variable's value range. Hence, second order statistics and least-mean-squares estimates that are sensitive to nonnormal degredations are not suitable, and we use instead the so-called nonparametric, robust statistical techniques and distance measures [15, 27, 14]. Out of the simplest robust location estimates, median and spatial median, we use spatial median due to it's multidimensional nature which allows better utilization of the local/clusterwise available data pattern [17]. Spatial median has many attractive statistical properties and, especially, it's breakdown point is 0.5, i.e. it can handle up to 50% of contaminated data.

In [4], a robust approach utilizing the spatial median to cluster sparse and noisy data was introduced. The *k-spatial-medians* clustering algorithm is based on the algorithmic skeleton as presented in Algorithm 1. As the score function one utilizes

$$\mathcal{J} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \|\boldsymbol{P}_i(\boldsymbol{x}_i - \boldsymbol{c}_j)\|_2, \tag{1}$$

where the last sum is computed over the subset of data attached to cluster j. Here the projections $\boldsymbol{P}_i, i = 1, \ldots, N$, capture the existing variable values of the ith observation, i.e.

$$(\boldsymbol{P}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases}$$

In Algorithm 1, the projected distance as defined in (1) is used in the first step, and recomputation of the prototypes, as spatial median with the available data, is realized using the SOR (Sequential Overrelaxation) algorithm [4] with the overrelaxation parameter $\omega = 1.5$.

3.1 Initialization and Number of Clusters

It is a well-known problem that all iterative clustering algorithms are highly sensitive to the initial placement of the cluster prototypes and, thus, such algorithms do not guarantee unique clustering [18, 9, 6, 16]. Numerous methods have been introduced to address this problem. Random initialization is still often chosen as the general strategy [31]. However, several researchers (e.g., [3, 5]) report that having some other than random strategy for the initialization often improves final clustering results significantly. Having these issues in mind, we developed the following deterministic and context-sensitive approach to find good initial prototypes.

For a subset of 2520 students in the Finnish data, there are

Algorithm 1: Iterative relocation clustering algorithm

```
Input: Dataset X with n observations and a given number
       of clusters k.
Output: A set of k clusters, which minimizes the score
        function.
Select k points as the initial prototypes;
repeat
    1. Assign individual observation to the closest
    prototype;
    2. Recompute the prototypes with the assigned
    observations;
until The partition does not change;
```

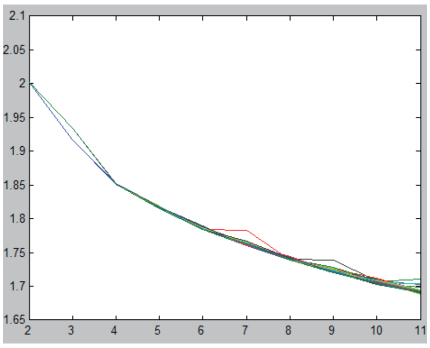


Figure 1: Ray-Turi index for $k = 2, \dots, 11$

no missing scale index values. For this subset we want to find (i) the most suitable amount of clusters k and (ii) the initial prototypes for the clustering algorithm with the whole data. For this purpose, we utilize a simple search strategy with two nested loops. The first loop iterates through different values of k and the second loop repeats the *k-spatialmedians* algorithm with random initialization ten times. For each clustering result, we then compute the so-called Ray-Turi index, see [25]. This index captures the principal purpose of clustering prototypes, i.e. accurate presentation of separate subset of data, and it is computed by simply dividing the score function (1) with the distance of the two closest prototypes. Figure 1 visualizes the plot of the Ray-Turi index for a set of values for the number of clusters. From the visualization we observe that the clustering result (Ray-Turi index) is decreasing when more clusters are introduced. However, after four clusters the speed of improvement is decreased. Moreover, for four clusters the result is very stable because all the ten random repetitions provide exactly the same clusters and prototypes. To this end, based on these observations, $k = 4$ is used as the number of clusters and the unique result for the full data as initialization for the whole, sparse data set clustering with Algorithm 1. The obtained result, characterized by four prototypes with available value for all scale indices, is to be interpreted next.

3.2 Interpretation of Clustering Result

The four cluster prototypes are depicted in Figure 2. Table 2 provides information about the students in the different clusters. Hereby, *valid indices* shows the percentage of existing index values in each cluster. As can be seen, the available data is quite evenly distributed among the clusters. While *sample size* denotes the actual number of students in the data, *population size of target group* is the same but each student is weighted so that they represent the whole Finnish population of 15-year-old students. *WA math score* is the weighted average of the mathematics scores from the students in the respective cluster.

As can be inferred from Figure 2 in combination with Table 2, we have one clear ”high performance” and one clear ”low performance” national cluster: The students in *Cluster 1* have mean performance in mathematics of 571.53 and they are on average the most advantaged students with highest beliefs in themselves. In all indices that are associated with highperformance in mathematics, the prototype that represents this cluster has the highest value. Solely in the ”intentions” to use mathematics later in their life, the students in *Cluster 1* lack behind the students in *Cluster 3*. *Cluster 4*, on the other hand, represents the most disadvantaged students in Finland, with lowest mean score in mathematics, and also lowest beliefs in themselves.

Cluster 2 and *Cluster 3* are, at the same time, similar and very different. According to the average performance of the students in those two clusters, both belong to PISA score Level 3 (see Table 4). As specified in the proficiency level descriptions in [22] this means that students in both of these clusters are able to, for example, solve tasks with clearly described procedures, but are unlikely to be able to (this proficiency is attributed to students from higher levels) also solve tasks that involve constraints or call for making assumptions. However, the prototypes (see Figure 2) show that students from these clusters can be opposite to each other by means of many scale indices.

While the students in *Cluster 2* generally are slightly more socially and economically advantaged, feel that they belong to school, and commonly have very positive attitude towards school, they definitely have below OECD average intentions to use mathematics, so that they also score worse in mathematics. *Cluster 2* is predominantly populated by girls. *Cluster 3*, on the other hand, has the lowest percentage of girls in it. This cluster consists of mostly boys who do not have the best attitude towards school. They also do not feel like they belong to school and generally are socially and economically less advantaged than the students in *Clusters 1* and *2*. However, they have the highest intentions to use mathematics later in their life, and pursue mathematics-related studies or careers in the future. They also tend to attribute failure in mathematics more to external factors than to themselves, have less anxiety towards mathematics than the OECD average, and are (although they do not seem to be interested in school in general) more interested in mathematics than the OECD average. It seems that they have already decided to have a career in a mathematics related profession, on the contrary to the (mostly female) students in *Cluster 2*.

As for the correlations before, we also created a ranking of indices to clarify the interpretation of the clustering result. The distance that defines the ranking to distinguish *Clusters 2* and *3* is just the absolute difference between the

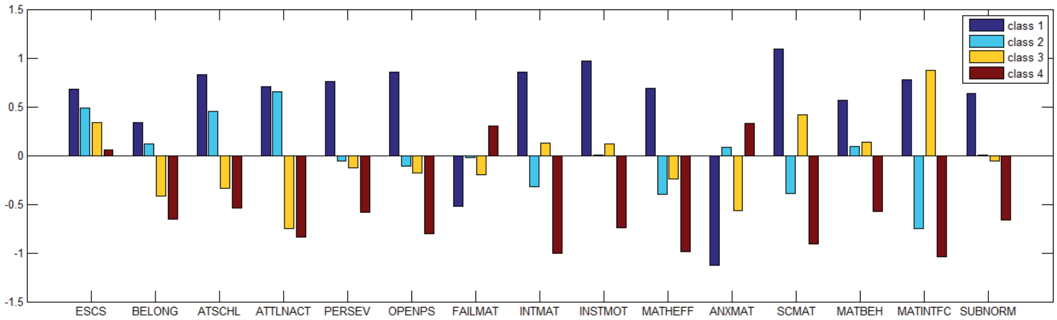


Figure 2: Clustering results

Table 2: Facts of clusters

	valid	sample	**population size of target group**			**WA math score**		
cluster	indices	size	all	♀ (in %)	♂	∅	♀	♂
C1	64%	1967	12884	5302 (41%)	7582	571.53	578.66	566.55
C2	69%	2192	14038	8598 (61%)	5440	509.82	516.76	498.85
C3	67%	2450	16751	6434 (38%)	10317	536.02	541.74	532.45
C4	66%	2220	16374	8876 (54%)	7498	467.21	472.96	460.40
C1-C4	67%	8829	60047	29210 (49%)	30837	518.75	520.19	517.39

Table 3: Separation of clusters

	all clusters		**Cluster 2 -3**	
index	distance	rank	distance	rank
ESCS	0.62	15	0.15	10
BELONG	0.98	13	0.53	6
ATSCHL	1.38	9	0.78	4
ATTLNACT	1.54	7	1.40	2
PERSEV	1.35	10	0.07	13
OPENPS	1.66	6	0.08	12
FAILMAT	0.83	14	0.17	8
INTMAT	1.86	3	0.44	7
INSTMOT	1.71	4	0.11	11
MATHEFF	1.68	5	0.16	9
ANXMAT	1.46	8	0.65	5
SCMAT	2.00	1	0.81	3
MATBEH	1.14	12	0.04	15
MATINTFC	1.91	2	1.63	1
SUBNORM	1.30	11	0.06	14

index values of the two prototypes. This is generalized as the distance between *all clusters* by simply summing the three absolute differences between individually ordered prototype indices. These two distances and the implied rankings are provided in Table 3. As can be seen from Table 3, the students' self-concept in mathematics, the index which also correlates the most with the performance in mathematics (see Table 1), discriminates all the clusters the most. It seems that students' beliefs in their own mathematics abilities capture their true knowledge and skills fairly well. Additionally, the intentions to use mathematics and the interest in this subject provide a good separation of the four clusters. Those two indices describe the students' drive and interest to learn mathematics because they perceive this subject as profitable and appealing to their future. The two interesting clusters, *Cluster 2* and *Cluster 3*, are separated the most by the intentions to pursue a career in mathematics and by the attitudes towards school concerning learning activities.

4. ASSOCIATION RULE DISCOVERY

The goal of association rule mining, one of the most utilized methods in EDM according to [8, 26], is to automatically find patterns that describe strongly associated attributes in data. The discovered patterns are usually represented in the form of implication rules or attribute subsets [1, 32]. We have two explicit clusters - *Cluster 1* which consists of the highest performing students and *Cluster 4* which consists of the lowest performing students - but for the two remaining clusters with mixed profile, *Cluster 2* and *Cluster 3*, we want to find patterns/rules that further characterize these students. Hence, we form for each student that belongs to one of these two clusters an itemset which contains the gender of the student (first subset in Table 4), all the scale indices (central subset in Table 4), and the categorized proficiency level in mathematics (last subset in this table).

PISA score levels define the performance level of the students. For example, for PISA 2012 the range of difficulty of tasks generates six levels of mathematics proficiency. Students with a performance score within the range of Level 1 are likely to be able to successfully complete Level 1 tasks, but are unlikely to be able to complete tasks at higher levels. Level 6 reflects tasks that are the most difficult in terms of mathematical skills and knowledge [22]. On average, both student clusters of interest belong to performance Level 3 (see Table 2). Therefore, in the corresponding item, we only distinguish three categories: below, within, or above Level 3 (see the last subset in Table 4).

Table 4: Items for Association Rules

id	item
1	girl
2	boy
3 & 4	$(+,-)$ ESCS
5 & 6	$(+,-)$ BELONG
7 & 8	$(+,-)$ ATSCHL
9 & 10	$(+,-)$ ATTLNACT
11 & 12	$(+,-)$ PERSEV
13 & 14	$(+,-)$ OPENPS
15 & 16	$(+,-)$ FAILMAT
17 & 18	$(+,-)$ INTMAT
19 & 20	$(+,-)$ INSTMOT
21 & 22	$(+,-)$ MATHEFF
23 & 24	$(+,-)$ ANXMAT
25 & 26	$(+,-)$ SCMAT
27 & 28	$(+,-)$ MATBEH
29 & 30	$(+,-)$ MATINTFC
31 & 32	$(+,-)$ SUBNORM
33	Level 2 or below: ≤ 482.38
34	Level 3: $482.38 - 544.68$
35	Level 4 or above: ≥ 544.68

In order to separate an individual student from main bulk of students, we fix a threshold value of 0.2 to define whether an item is part of the itemset for that particular student. The threshold 0.2 is chosen because it provides the median (rounded to one decimal place) of the absolute values of scale indices of all cluster prototypes. If a positive index value for a certain student is above the threshold, then the first *id* in the matrix (see Table 4) will be part of the itemset. Similarly, if a negative index value is below the negative threshold, then the second *id* (see Table 4) will belong to the itemset. Again, we utilize only the available indices. This means that in case the student's index value is inside $[-0.2, 0.2]$ or missing/invalid, it is not included in the itemset. For finding frequent itemsets based on the encoding, we used the implementation described in [13], and for generating association rules from the obtained frequent itemsets we utilized the implementation explained in [2].

4.1 Basic Concepts of Frequent Itemsets

Let I be the set of all items. An important property of an itemset is its *support count*, which refers to the number of transactions that contain a particular itemset. Let S_1 be a subset of the set of items ($S_1 \subseteq I$). Logically, a transaction $t_i \in T$, where T denotes the set of all transactions, is said to contain itemset S_1 if S_1 is a subset of t_i. Mathematically, the support count, $\sigma(S_1)$, for an itemset S_1 can be stated as follows:

$$\sigma(S_1) = |\{t_i \mid S_1 \subseteq t_i, t_i \in T\}|,$$

where $|\cdot|$ stands for the number of elements in a set. An *Association Rule* is then an implication expression of the form $S_1 \to S_2$, where $S_1, S_2 \subseteq I$ and $S_1 \cap S_2 = \emptyset$.

The support, $s(S_1 \to S_2)$, determines how often a rule is applicable to a given data set. Furthermore, the confidence, $c(S_1 \to S_2)$, determines how frequently items in S_2 appear in the transactions that contain S_1. Mathematically this can be expressed as follows:

$$s(S_1 \to S_2) = \frac{\sigma(S_1 \cup S_2)}{|T|} \text{ and } c(S_1 \to S_2) = \frac{\sigma(S_1 \cup S_2)}{\sigma(S_1)},$$

Support measures how well a rule is covered by the data. Therefore, if a rule has a too low support, it could be that it occurred solely by chance. Confidence is an important measures as it provides the the reliability and accuracy of a rule.

4.2 Obtained Rules and Interpretation

When we use the applied implementation of the famous Apriori Algorithm, we obtain many trivial rules. For example, it is already obvious from the clustering prototypes that those students who have highly positive attitude towards learning activities have also highly positive attitude towards learning outcomes. However, as already discussed, our itemsets can be divided into three subsets: the set that contains the gender, the set which contains the performance in mathematics, and the set which contains the different scale indices. We are interested in the gender differences and the performance in mathematics. Therefore, we search inside the algorithm's output for rules that have items of the gender and/orperformance interval subsets at the right hand side of the rule.

We start with high values for support and confidence and lower then the confidence threshold. Since we are especially interested in rules that contain the gender, the support has to have a relatively small value, so we choose the minimum value 0.1 while trying to keep the confidence value as high as possible. Starting with confidence of 1 and lowering it successively, we obtain the first rule that has gender on the right side with confidence 0.71:

$$\{\text{-ATTLNACT, +SCMAT, +MATINTFC}\} \Rightarrow \{boy\} \tag{2}$$

In words (2) means that those students who have negative attitudes towards school but a high self-concept and high intentions in mathematics are boys.

The first rule that we obtain for girls with confidence 0.69 is of the form:

$$\{\text{ -MATHEFF, - MATINTFC}\} \Rightarrow \{girl\} \tag{3}$$

Rule (3) says that those students who have negative self-efficacy and no intention to use mathematics are girls.

If we lower the minimal acceptable support into 0.095, we obtain the following interesting rule (4): Those students who have positive attitudes towards school but no intention to use mathematics later in life are girls.

$$\{\text{+ATTLNACT, -MATINTFC}\} \Rightarrow \{girl\} \tag{4}$$

Next, with the same minimal support we are searching explicitly for rules that have performance value below or above Level 3 at the left-hand side of the rule and gender at the right-hand side. Here, we first obtain the following rule with a confidence value of 0.6:

$$\{\text{+ATTLNACT, above Level 3 performance}\} \Rightarrow \{girl\} \tag{5}$$

According to (5), those students with a proficiency level above 3 and a clearly above average positive attitude towards learning activities in school are girls.

With confidence 0.52 we obtain the first rule for boys:

$$\{\text{+SCMAT, above Level 3 performance}\} \Rightarrow \{boy\} \tag{6}$$

Rule (6) means that those students with a proficiency level above 3 and a clear above average self-concept in mathematics are boys.

Subsequently, we are searching for rules wich have both gender and below or above Level 3 performance at the left-hand side of the rule. Such rule with the highest confidence (0.65) reads as:

$$\begin{aligned}&\{\text{-ATSCHL -ATTLNACT +OPENPS -FAILMAT}\\&\quad\text{+SCMAT}\} \Rightarrow \{boy, \text{above Level 3 performance}\}\end{aligned} \tag{7}$$

According to (7), those students with negative attitudes towards school (both, learning outcome as well as learning activities) but with clearly above average openness to problem solving, a high self-concept in mathematics and strictly below average self-responsibility for failing in mathematics, are boys that perform above Level 3.

For girls the rule with the highest confidence (0.63) is given by (8):

$$\begin{aligned}&\{\text{-ESCS +ATTLNACT +ANXMAT -SCMAT}\}\\&\qquad\Rightarrow \{girl, \text{below Level 3 performance}\}\end{aligned} \tag{8}$$

This means that those students who are socially and economically less advantaged, have high anxiety towards mathematics and a low self-concept in mathematics, but still clearly above average attitude towards school, are girls who perform below Level 3.

If we unite the rules given in (2)-(8), we see that in all the rules that contain boys the item which represents the high self-concept in mathematics is present. In general, high-performing boys are also convinced that they can succeed (see 6). Moreover, even when they fail in mathematics, they are more likely to see other individuals or factors responsible on this than themselves (see 7). In addition, they have the highest intentions to use mathematics later in their life (see 2). However, according to the rules, male students can have negative attitude towards school (see 2 and 8), whereas the most positive attitudes appear only in the rules that include girls. Even the below average performing and socially and economically more disadvantaged girls with low self-concept and high anxiety towards mathematics, perceive the learning activities in their schools as very important (see 8). The same positive attitude towards school is also associated with the highest performing girls (see 5). Moreover, female students are much less confident about their mathematic skills (see 3) and have least intentions to pursue a mathematics related career (see 3 and 4).

To sum up, we conclude that specific characteristics and attitudes in the two middle performing clusters are, indeed, often gender-specific. Since we explicitly searched for rules that have certain items in them, we can not express precisely how typical these situations are. Nevertheless, when we combine all obtained rules with the clustering result two main characterizations appear: On the one hand, we have a specific subgroup of mainly girls who we nominated ”to-be-nurses”: they seem to be capable of performing well if they want to, having strongly positive attitude towards school. However, these students have low beliefs in themselves to be able to succeed in mathematics, and even a somewhat fear towards mathematics. On the other hand, we have a subgroup of mainly boys which we refer as ”to-be-engineers”. These students do not seem very interested in school in general. Yet, they trust in their capabilities and are extremely confident about their skills to perform well in mathematics. Even if they fail, they attribute this failure rather to other external factors than to themselves.

5. SUMMARY AND CONCLUSIONS

Although Finland is one of the few countries in which, on average, girls perform slightly better than boys in mathematics, professional careers related to this subject are also in here still dominated by men. We have applied methods from two of the main educational data mining branches on PISA data to obtain more gender-specific knowledge which might explain this observation.

First of all, we utilized a special robust clustering approach to group the students according to those PISA scale indices that are associated with performance in mathematics. The index that represents the students' self-concept in mathematics (SCMAT), and which also was the variable that correlates the most with the students' performance in mathematics (see Table 1), is the most important discriminator for the four clusters that we obtained (see Table 3). Combined with the other attributes we conclude that those students who have a higher self-concept, and tend to be socially and economically more advantaged, perform better than their less advantaged peers. They also have better attitudes to school, trust more in their own capabilities, and have greater expectation for their future careers (see Figure 2).

Two of the clusters we obtained, *Cluster 1* representing the ”high performing” and *Cluster 4* representing the ”low performing” students, can to a large extend be explained by these differences. However, the two ”medium” clusters show the opposite behaviour: Socially and economical more advantaged students with very positive attitudes towards school and learning from *Cluster 2* perform worse in mathematics than the somewhat more disadvantaged students in *Cluster 3*. We found that these clusters are separated the most by the index that measures the student's intentions to pursue a mathematics related career. Since *Cluster 2* is with 61% dominated by girls, while *Cluster 3* consists of a larger percentage (62%) of boys we assumed that this difference could be explained by the gender of the student.

Association rule mining in the data subset of these two remaining medium clusters revised the gender-specific attitudes even more, and confirmed our assumption. Those 15-year-old students from this subset who already seem to have decided to pursue a mathematics related career are mostly boys. On the other hand, the attribute that is the most ascribable to girls is the positive attitude towards school. Altogether, the results of our study suggest that there are distinct groups of high and low performing students. However, the bulk of the girls with average performance seem to have no intentions to pursue a mathematics related profession. This is neither connected to their social status nor to their attitudes towards school. In fact, they often show a better feeling of belonging to school and have very positive attitudes towards school and learning. While boys often consider mathematics as a great part of their future even when they do not show obvious skills, girls tend to be discouraged much faster and to easier favour other subjects. We feel that this is an important finding that should be studied further, especially concerning when such a gender-specific orientation starts to emerge.

6. REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.

[2] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[3] R. T. Aldahdooh and W. Ashour. DIMK-means ”Distance-based initialization method for K-means clustering algorithm”. *International Journal of Intelligent Systems and Applications (IJISA)*, 5(2):41, 2013.

[4] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[5] L. Bai, J. Liang, and C. Dang. An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems*, 24(6):785–795, 2011.

[6] L. Bai, J. Liang, C. Dang, and F. Cao. A cluster centers initialization method for clustering categorical data. *Expert Systems with Applications*, 39(9):8022–8029, 2012.

[7] R. Baker et al. Data mining for education. *International Encyclopedia of Education*, 7:112–118, 2010.

[8] R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1, 2010.

[9] M. Emre Celebi, H. A. Kingravi, and P. A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 2012.

[10] V. Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1):65–75, 2002.

[11] A. L. Goodwin. Perspectives on high performing education systems in Finland, Hong Kong, China, South Korea and Singapore: What lessons for the US? In *Educational Policy Innovations*, pages 185–199. Springer, 2014.

[12] J. Han, M. Kamber, and A. Tung. Spatial clustering methods in data mining: A survey, 2001.

[13] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[14] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Edward Arnold, London, 1998.

[15] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., New York, 1981.

[16] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[17] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[18] M. Meilă and D. Heckerman. An experimental comparison of several clustering and initialization methods. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 386–395. Morgan Kaufmann Publishers Inc., 1998.

[19] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[20] OECD. *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)*. PISA, OECD Publishing, 2013.

[21] OECD. *PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III)*. PISA, OECD Publishing, 2013.

[22] OECD. *What Makes Schools Successful? Resources, Policies and Practices (Volume IV)*. PISA, OECD Publishing, 2013.

[23] OECD. *PISA 2012 Results: What Students Know and Can Do. Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. PISA, OECD Publishing, 2014.

[24] N. Raheja and R. Kumar. Optimization of association rule learning in distributed database using clustering technique. *International Journal on Computer Science & Engineering*, 4(12), 2012.

[25] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[26] C. Romero and S. Ventura. Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618, 2010.

[27] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons Inc., New York, 1987.

[28] M. Saari. Promoting gender equality without a gender perspective: Problem representations of equal pay in Finland. *Gender, Work & Organization*, 20(1):36–55, 2013.

[29] M. Steinbach, L. Ertöz, and V. Kumar. The challenges of clustering high dimensional data. In *New Directions in Statistical Physics*, pages 273–309. Springer, 2004.

[30] B. Xu, M. Recker, X. Qi, N. Flann, and L. Ye. Clustering educational digital library usage data: A comparison of latent class analysis and k-means algorithms. *Journal of Educational Data Mining*, 5(2):38–68, 2013.

[31] R. Xu and D. C. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[32] Q. Zhao and S. S. Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 2003.

PIV

WEIGHTED CLUSTERING OF SPARSE EDUCATIONAL DATA

by

Mirka Saarela, Tommi Kärkkäinen 2015

Proc. of the 23rd European Symposium on Artificial Neural Networks,
Computational Intelligence and Machine Learning, pp. 337–342

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Weighted Clustering of Sparse Educational Data

Mirka Saarela and Tommi Kärkkäinen

University of Jyväskylä - Department of Mathematical Information Technology
40014, Jyväskylä - Finland

Abstract. Clustering as an unsupervised technique is predominantly used in unweighted settings. In this paper, we present an efficient version of a robust clustering algorithm for sparse educational data that takes the weights, aligning a sample with the corresponding population, into account. The algorithm is utilized to divide the Finnish student population of PISA 2012 (the latest data from the Programme for International Student Assessment) into groups, according to their attitudes and perceptions towards mathematics, for which one third of the data is missing. Furthermore, necessary modifications of three cluster indices to reveal an appropriate number of groups are proposed and demonstrated.

1 Introduction

The application of clustering in a weighted context is a relatively unresearched topic [1]. PISA (Programme for International Student Assessment) is a worldwide study that triannually assesses proficiency of 15-year-old students from different countries and economies in the three domains, reading, mathematics, and science. Besides the reporting of student performances, PISA is also one of the largest public databases[1] in which students' demographic and contextual data, such as their attitudes and behaviors towards education related topics, is collected and stored.

PISA data are an important example of a large data set that includes weights. In general, weighting is a technique in survey research to align the sample to more accurately represent the true population. Namely, only a fraction of students from each country take part in the PISA assessment but, when taking the weights into account, they should be representative for the whole population. For example, the Finnish sample data of the latest PISA assessment consists of 8829 students whose analysis results, when multiplied with the respective weights, represent the whole 60047 15-year-old student population of the country. As can be seen from Fig. 1, in which the studentwise weights are depicted, the minimal weight in the Finnish national subset of PISA is 1, i.e. each students represents at least him/herself, while the maximal weight is more than 54.

A further important characteristic of PISA data is the large number of missing values. Because PISA uses a rotated design [2] and some students are not administered certain questions, the majority of the missing data in PISA is missing by design, which can be seen as a special case of *missing completely at random* [3, 4]. Altogether, there are 634 raw variables in the PISA student questionnaire data set of the latest assessment. However, a subset of 15 derived

[1] PISA data can be downloaded from `http://www.oecd.org/pisa/pisaproducts/`.

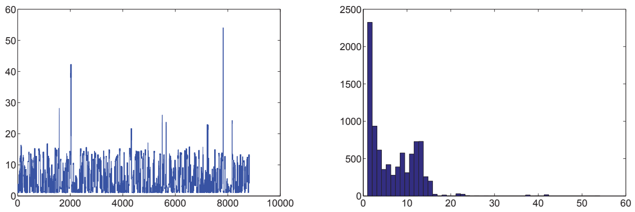


Fig. 1: Individual weights (left) and their discrete distribution (right) in Finnish 2012 PISA data.

variables, the so-called PISA scale indices[2], readily describe students' attitudes and perceptions, e.g., explaining the performance in mathematics [2, 5]. Each scale index is a compound variable and constructed using the students' answers to certain background questions. Nevertheless, mainly because of the rotated design, 33.24% of these scale indices are not available.

In [5] we utilized a robust clustering algorithm to the Finnish sample of PISA 2012 scale indices, which revealed very gender-specific contrasts in the different clusters. For the interpretation of the clustering result, we employed the weights to summarize the cluster prototypes on the population level. However, according to the PISA data analysis manual [6], one should always, particularly when over- or under-sampling has taken place, include weights at *each stage* of the analysis.

Therefore, the research questions of this paper are as follows: (i) how to efficiently cluster sparse student data on the population level, i.e., how the weights in the sample should be incorporated in the robust clustering algorithm and (ii) how much the two clustering results with and without weights (sample division vs. population division) differ from each other? Both questions are relevant for the Finnish subset of PISA data because immigrants as well as students from Swedish-speaking schools were deliberately over-sampled in the latest assessment.

2 Weighted robust clustering of sparse data

In general, partitioning-based clustering algorithms are composed of an initialization followed by the iterations of the two basic steps, where each observation is first assigned to its closest prototype and, then, each prototype is updated based on the assigned subset of data. As pointed out in [5], sparse data sets can be reliably clustered by utilizing the so-called *k-spatialmedians* [7] algorithm. Compared to k-means, the k-spatialmedians uses the spatial median to estimate the prototypes, which is statistically robust and can handle large amount of contamination (noise and missing values) in data.

However, because of the local search character of the partitioning-based clustering algorithms, their result depends on the initialization. For a sparse data set

[2] These scale indices are explicitly listed in [5].

with missing values, a proper initialization should posses, at least, two desired properties: it should reflect the subset of data with full observations, because inevitably missing values decrease reliability of the cluster allocations. Furthermore, the initial prototypes should be full, i.e., without missing values, because the cluster assignment and recomputation, e.g., as in [5], assumes this throughout the whole iterative procedure. Lately the k-means++ algorithm [8], where the random initialization is based on using a density function favoring distinct prototypes, has become popular.

Therefore, our general procedure to cluster the sparse data on the population-level is as follows. First of all, the subset of data that has no missing values is clustered using k-means++. Then, the robust clustering algorithm is applied for the whole sparse data by utilizing the obtained prototypes as initialization. Altogether, the final clustering result is statistically robust with respect to degradations in data, probably with full prototypes (especially when a small number of clusters is created from a large data set), and reflecting the spherical and possibly already separated shape of the full data subset.

The precise form of the general clustering criterion to be minimized (locally) by the iterative reallocation algorithm, with weights and missing values, reads as follows:

$$\mathcal{J}(\{c_k\}_{k=1}^K) = \sum_{k=1}^K \sum_{i \in I_k} w_i \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2^p, \tag{1}$$

where I_k denotes the indices of data assigned to the kth cluster and $\boldsymbol{P}_i$'s define the sparsity pattern (i.e., indicate available variables) observationwise:

$$(\boldsymbol{P}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases}$$

In the k-spatialmedians algorithm for $p = 1$, the cluster prototypes are computed using a modifed SOR (Sequential Overrelaxation) algorithm [7], where weights are taken into account in the updates. Furthermore, in order the align the k-means-type initialization with $p = 2$ in (1) to the actual case $p = 1$, we propose to use $\{\sqrt{w_i}\}$'s as weights in k-means++ because, simply, $\alpha \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2^p = (\sqrt[p]{\alpha} \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2)^p, \quad \text{for } \alpha > 0.$

To this end, to determine a single result of the partitioning-based weighted clustering procedure, one also needs to estimate the number of clusters K. For this purpose, we used three modified internal cluster validation indices, namely the Ray-Turi [9], the Davies-Bouldin [10], and the Davies-Bouldin* [10]. Essentially, we included the weights in the computations of the clusterwise scatter matrices, used the final value of (1) as the clustering error, and computed distances between the prototypes by using the Euclidean norm.

3 Experimental results

The tests concentrate on analyzing the use of weights in the initial partition utilizing k-means++, followed by the actual weighted k-spatialmedians. Namely,

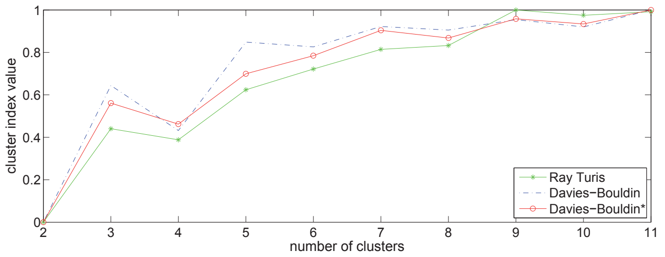


Fig. 2: Cluster indices for sparse data scaled into range $[0, 1]$.

one can use/omit the weights in i) the initialization of k-means++ and ii) the iterative reallocations of k-means++, which creates three possible algorithmic scenarios. First of all, all of these possibilities were applied to assess the number of clusters using the modified cluster indices. The result is given in Fig. 2 where the averages of 30 runs (ten for each variant for each k) is depicted. One concludes that all three cluster indices suggest that, for the Finnish 2012 population data, four clusters is an appropriate choice[3]. This is the same number that was obtained for the Finnish sample data without weighting (see [5]).

Next we fix $k = 4$, i.e., test the speed (number of iterations) and quality of the three algorithmic combinations for four clusters. The results with 10 repeated test runs are given in Table 1, together with the average of the ten repetitions in the last row. We report the number of iterations needed in the initialization (i.e. within k-means++), the number of iterations needed in the actual k-spatialmedians clustering with the whole sparse data, and also the final quality of the clustering result (i.e., the clustering error).

All three main columns of Table 1 show that including the weights in k-means++ for complete data before k-spatialmedians improves the performance of the latter as less iterations are needed. Similarly, to include square-rooted weights[4] in the initialization of k-means++ improves the performance of the whole initial procedure (see the last two main columns). Concerning the clustering error, we obtained similar error levels with all the approaches (see the last row of Table 1) but less variability when using the weights. Therefore, we conclude that appropriately scaled weights should be present in both places in the initialization in order to achieve an efficient and robust weighted clustering algorithm.

Using the fully weighted algorithm with the average of 10 runs, we obtain in practice the same four clusters as in the unweighted case (see [5] in which the clusters and their implications are discussed) with very similar characteristics

[3] Actually, all three indices have the best value at two but having only two clusters divides our data simply in high- and low-performing students which does not provide any interesting patterns additionally.

[4] Incorporating the weights into k-means++ simply as w instead of $\sqrt{w}$ was also tested. But since $\sqrt{w}$ gave, as we proposed in Sec. 2, better results, only these are reported here.

(see Table 2). The prototypes that describe the four clusters are almost identical. In particular, also with weights the cluster *C2* of mostly girls, with very positive attitudes towards school and learning but no intentions to use mathematics later in life, appear. Also an opposite cluster *C3* with the majority of boys, that have the highest intentions to pursue a mathematics related career but otherwise very negative attitudes towards education, is present, together with the groups of advantaged high-performing students (*C1*) and their more disadvantaged lower performing peers (*C4*).

4 Conclusions

In this paper, we modified the k-spatialmedians algorithm [7], an algorithm that can handle large amounts of missing data, in such a way that it can be used also for weighted clustering. In order to have an as fast and deterministic approach as possible, we also introduced weights to the seeding as well as the actual main body of the k-means++ algorithm which we use in the initialization. Experiments showed that, indeed, the best, i.e. the fastest as well as most accurate, population-based clustering solution is obtained when weights are incorporated in all phases of the algorithm.

As pointed out in the introduction, though weighted clustering has been investigated in theory, it has not been examined much in an applied context. PISA data sets are prime examples of large data sets with many missing values as well as weights. We applied weighted clustering to the Finnish subset of the latest PISA data. Although over-sampling took place for some groups of the student population, no significant differences in the final results existed, i.e. the general

Without weights in k-means++			$\sqrt[p]{w_i}$ weights in iterative reallocation			$\sqrt[p]{w_i}$ weights in entire algorithm		
iter. in ini.	iter. in alg.	cluster error (quality)	iter. in ini.	iter. in alg.	cluster error (quality)	iter. in ini.	iter. in alg.	cluster error (quality)
23	34	5.9464	34	30	0.6458	21	28	0.6035
23	38	0.5176	34	30	0.6458	14	30	0.5424
19	33	0.5161	41	33	0.5176	23	30	0.5424
27	38	0.5176	42	30	0.5176	29	30	0.5424
23	34	0.4983	34	33	0.6458	18	29	0.5424
23	38	0.5176	34	30	0.6458	20	30	0.5424
21	44	6.0403	43	30	0.6458	22	30	0.5424
18	38	0.5176	39	33	0.5176	24	30	0.5424
25	38	0.5176	41	33	0.6458	26	28	0.6035
20	37	0.5176	34	30	0.6458	22	28	0.6035
20	38	1.6108	41	31	0.6073	22	29	0.5607

Table 1: Efficacy and quality of clustering result with and without weights in initialization. The base level 127450 has been subtracted from all cluster errors.

	valid	sample	**population size**			**math score**		
cluster	indices	size	all	♀ (in %)	♂	∅	♀	♂
C1	65%	2009	13203	5311 (40%)	7893	574	581	569
C2	68%	2242	14418	8955 (62%)	5463	510	516	499
C3	67%	2450	16723	6495 (39%)	10229	532	539	528
C4	66%	2128	15703	8450 (54%)	7253	466	472	460
C1-C4	67%	8829	60047	29210 (49%)	30837	519	520	517

Table 2: Facts of population clusters

profiles of the clusters without weights (sample) and with weights (population) were almost identical. However, even though the algorithm is deterministic after the initialization, and the accuracy of clustering is improved when initialized with k-means++, still some randomness in the final clustering result remains due to the randomness in seeding. Hence, a complete comparison between clustering results persists challenging, not only for population- vs. sample-based clustering but also for clustering in general.

References

[1] Margareta Ackerman, Shai Ben-David, Simina Branzei, and David Loker. Weighted clustering. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[2] OECD. Pisa 2012 technical background. 2013.

[3] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[4] Donald B Rubin and Roderick JA Little. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*, 2002.

[5] Mirka Saarela and Tommi Kärkkäinen. Discovering gender-specific knowledge from finnish basic education using pisa scale indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[6] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[7] Sami Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[8] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[9] Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[10] Minho Kim and RS Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

PV

DO COUNTRY STEREOTYPES EXIST IN PISA? A CLUSTERING APPROACH FOR LARGE, SPARSE, AND WEIGHTED DATA

by

Mirka Saarela, Tommi Kärkkäinen 2015

Proc. of the 8th International Conference on Educational Data Mining, pp. 156–163

Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data.

Mirka Saarela
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
mirka.saarela@gmail.com

Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
tommi.karkkainen@jyu.fi

ABSTRACT

Certain stereotypes can be associated with people from different countries. For example, the Italians are expected to be emotional, the Germans functional, and the Chinese hard-working. In this study, we cluster all 15-year-old students representing the 68 different nations and territories that participated in the latest Programme for International Student Assessment (PISA 2012). The hypothesis is that the students will start to form their own country groups when clustered according to the scale indices that summarize many of the students' characteristics. In order to meet PISA data analysis requirements, we use a novel combination of our previously published algorithmic components to realize a weighted sparse data clustering approach. This enables us to work with around half a million observations with large number of missing values, which represent the population of more than 24 million students globally. Three internal cluster indices suitable for sparse data are used to determine the number of clusters and the whole procedure is repeated recursively to end up with a set of clusters on three different refinement levels. The results show that our final clusters can indeed be explained by the actual student performance but only to a marginal degree by the country.

Keywords

Weighted Clustering, PISA, Sparse Cluster Indices, Country Stereotype

1. INTRODUCTION

Certain stereotypes seem to be associated with people from different countries. The French and Italians, for example, are expected to be emotional, while Germany has mainly a functional country stereotype [4], and the Chinese are commonly perceived as hard-working [3]. According to the *Hofstede Model* [6], national cultures can be characterized along six dimensions: power distance, individualism, masculinity, uncertainty avoidance, pragmatism, and indulgence. The hypothesis in this study is that also the population of 15-year-old students worldwide will start to form their own national groups, i.e., show similar characteristics to their country peers, when clustered according to their attributes and attitudes towards education.

PISA (Programme for International Student Assessment) is a worldwide triannual survey conducted by the Organisation for Economic Co-operation and Development (OECD), assessing the proficiency of 15-year-old students from different countries and economies in three domains: reading, mathematics, and science. Besides evaluating student performances, PISA is also one of the largest public databases[1] of students' demographic and contextual data, such as their attitudes and behaviours towards various aspects of education.

In order to test our hypothesis, we utilize the 15 PISA scale indices (explicitly detailed in [14]), a set of derived variables that readily summarize the background of the students including their characteristics and attitudes. In particular, the *escs* index measures the students' economic, social and cultural status and is known to account for most variance in performance [9]. Additionally, 5 scale indices (*belong*, *atschl*, *attlnact*, *persev*, *openps*) are generally associated with performance on a student-level, while 9 further ones (*failmat*, *intmat*, *instmot*, *matheff*, *anxmat*, *scmat*, *mathbeh*, *matintfc*, *subnorm*) are directly related to attitudes towards mathematics, the main assessment area in the most recent survey (PISA 2012). However, since the assessment material exceeds the time that is allocated for the test, each student is administered solely a fraction of the whole set of cognitive items and only one of the three background questionnaires. Because of this rotated design, 33.24% of the PISA scale indices values are missing.

Moreover, PISA data are an important example of large data sets that include weights. Only some students from each country are sampled for the study, but multiplied with their respective weights they should represent the whole 15-year-old student population. The sample data of the latest PISA assessment, i.e., the data we are working with, consists of 485490 students which, taking the weights into account, represent more than 24 million 15-year-old students in the 68 different territories that participated in PISA 2012.

[1] See `http://www.oecd.org/pisa/pisaproducts/`.

The content of this paper is as follows. First, we describe the clustering algorithm that allows us to work with the large, sparse and weighted data (Sec. 2). Second, we present the clustering results (Sec. 3) and their relevance to our hypothesis, i.e., how the clusters on the different levels can be characterized and to what extent they form their own country groups. Finally, in Sec. 4, we conclude our study and discuss directions for further research.

2. THE CLUSTERING APPROACH

Sparsity of PISA data must be taken into account when selecting or developing a data mining technique. With missing values one faces difficulties in justifying assumptions on data or error normality [14, 15], which underlie the classical second-order statistics. Hence, the data mining techniques here are based on the so-called nonparametric, robust statistics [5]. A robust, weighted clustering approach suitable for data sets with a large portion of missing values, non-normal error distribution, and given alignment between a sample and the population through weights, was introduced and tested in [16]. Here, we apply a similar method with slight modifications, along the lines of [7] for sampled initialization and [17] for hierarchical application. All computations were implemented and realized in Matlab R2014a.

2.1 Basic method

Denote by N the number of observations and by n the dimension of an observation of the data matrix $\mathbf{X}$; and let $\{w_i\}, i = 1, \dots, N$ be the positive sample-population-alignment weights. Further, let $\{\mathbf{p}_i\}, i = 1, \dots, N$, be the projection vectors that define the pattern of the available values [10, 1, 14, 15]. The weighted spatial median $\mathbf{s}$ with the so-called available data strategy can be obtained as the solution of the projected Weber problem

$$\min_{\mathbf{v}\in\mathbf{R}^n} \mathcal{J}(\mathbf{v}), \quad \mathcal{J}(\mathbf{v}) = \sum_{i=1}^{N} w_i \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{v})\|, \qquad (1)$$

where $\mathrm{Diag}\{\mathbf{p}_i\}$ denotes the diagonal matrix corresponding to the given vector $\mathbf{p}_i$. As described in [8], this optimization problem is nonsmooth, i.e., it is not classically differentiable. However, an accurate approximation for the solution of the nonsmooth problem can be obtained by solving the regularized equation (see [1]) $\sum_{i=1}^{N} \frac{w_i \mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\max\{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|,\delta\}} = \mathbf{0}$ for $\delta > 0$. This is solved using the SOR (Sequential Overrelaxation) algorithm [1] with the overrelaxation parameter $\omega = 1.5$. We choose $\delta = \sqrt{\varepsilon}$ for ε representing the machine precision.

In case of clustering with K prototypes, i.e., the centroids that represent the K clusters, one determines these by solving the nonsmooth problem $\min_{\{\mathbf{c}_k\}_{k=1}^{K}} \mathcal{J}(\{\mathbf{c}_k\})$, where all $\mathbf{c}_k \in \mathbf{R}^n$ and

$$\mathcal{J}(\{\mathbf{c}_k\}) = \sum_{k=1}^{K} \sum_{i\in I_k} w_i \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{c}_k)\|. \qquad (2)$$

Hereby, I_k determines the subset of data being closest to the kth prototype $\mathbf{c}_k$. The main body of the so-called iterative relocation algorithm for minimizing (2), which is referred as *weighted k-spatialmedians*, consists of successive application of the two main steps: i) find the closest prototype for each observation, and ii) recompute all prototypes $\mathbf{c}_k$ using the attached subset of data. For the latter part, we compute the weighted spatial median as described above. Note that the first step of finding the closest prototype of the ith observation, $\min_k \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{c}_k)\|$, does not need to take the positive weight w_i in (2) into account.

The next issues for the proposed method are the determination of the number of clusters K and the initialization of the clustering algorithm for a given k. Basically, the quality of a cluster can be defined by minimal within-cluster distances and maximal between-cluster distances. Therefore, for the first purpose, we use the approach suggested in [16] and apply three internal cluster indices, namely *Ray-Turi (RT)* [13], *Davies-Bouldin (DB)* [2], and *Davies-Bouldin* (DB*)* [11]. All these indices take both aspects of clustering quality into account: In essence, the clustering error (2), i.e., the sum of the within-cluster distances, to be as small as possible, is divided with the distance between the prototypes (minimum distance for RT and different variants of average distance for DB and DB*), to be as large as possible. When testing a number of possible numbers of prototypes from $k = 2$ into $K_{\max}$, we stop this enlargement when all three cluster indices start to increase.

Concerning the initialization, again partly similarly as in [16], we use a weighted k-means++ algorithm in the initialization of the spatial median based clustering with the weights $\sqrt{w_i}$. A rigorous argument for such an alignment was given in [9] where the relation between variance (weighted k-means) and standard deviation (weighted *k-spatialmedians*) was established. Because of local character, the initialization and the search are repeated $N_s = 10$ times and the solution corresponding to the smallest clustering error in (2) is selected. Furthermore, the weighted k-means++ is applied in the ten initializations with ten different, disjoint data samples (10% of the whole data) that were created using the so-called *Distribution Optimally Balanced, Stratified Folding* as proposed in [12], with the modified implementation given in [7]. Such sampling, by placing a random observation from class j and its $N_s - 1$ nearest class neighbors into different folds, is able to approximate both classwise densities and class frequencies in all the created data samples. Here, we use the 68 country labels as class indicators in stratification.

2.2 Hierarchical application

Because a prototype-based clustering algorithm always works with distances for the whole data, the detection of clusters of different size, especially hierarchically on different scales or levels of abstraction, can be challenging. This is illustrated with the whole PISA data set in Fig. 1, which shows the values of the three cluster indices for $k = 1, \dots, 68$. For illustration purposes, also the clustering error as defined in (2), denoted as ‘Elbow’, is provided. All indices have their minimum at $k = 2$ which suggest the division of the PISA data to only two clusters. Note that the geometrical density and low separability of the PISA scale indices might be related to their standardization to have zero mean and unit variance over the OECD countries.

Hierarchical application of the *k-spatialmedians* algorithm was suggested in [17]. The idea is simple: Similarly to the divisive clustering methods, apply the algorithm recursively

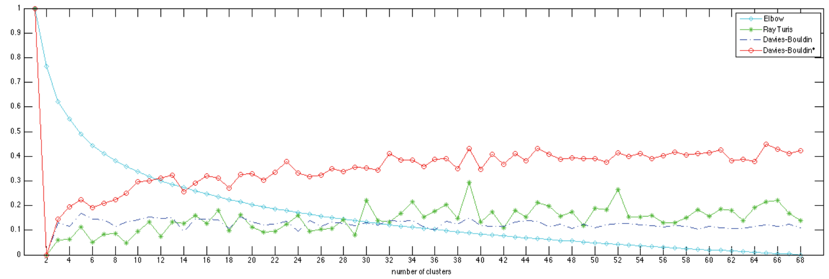


Figure 1: Cluster indices and error slope for the whole sparse PISA data scaled into range $[0, 1]$.

to the cluster data sets that have been determined using the basic approach. For the PISA data here, we realized a recursive search of the *weighted k-spatialmedians* with the depth of three levels, ending up altogether with 2 (level 1), $4+4$ (level 2), and $6+12+10+6$ & $2+8+3+6$ clusters (level 3). The wall-clock time for each individual clustering problem was several hours.

3. RESULTS

As discussed in Sec. 1, we use the 15 PISA scale indices that readily summarize most of the students' background as data input for our clustering algorithm. By following the mixture of the partitional/hierarchical clustering approach as described above, we first of all, provide the results of the weighted sparse data clustering algorithm when applied to the whole PISA data (first level). Then, recursively, the results of the algorithm for the newly obtained clusters at the second and third level of refinement are given. For all the clusters at each level, we compute the relative share of students from each country, i.e., the weighted number of students in the cluster in relation to the whole number of 15-year-old students in the country. Moreover, in order to reveal the deviating characteristics of the appearing clusters, we visualize and interpret (i.e., characterize) the cluster prototypes in comparison to the overall behavior of the entire 15-year-old student population in the 68 countries by always subtracting the weighted spatial median of the whole data from the obtained prototypes.

3.1 First Level

Since, as pointed out in Sec. 2.2, all the sparse cluster indices suggest two, we first run our weighted sparse clustering algorithm for $K = 2$. The clustering result on the first level is shown in Fig. 2. The division of these clusters is unambiguous: All scale indices that are associated with high performance in mathematics have a positive value for Cluster 2 and a negative value for Cluster 1. Likewise, those two scale indices that are associated with low performance in mathematics, i.e., the self-responsibility for failing in mathematics (*failmat*) and the anxiety towards mathematics (*anxmat*), show a positive value for Cluster 1 and a negative value for Cluster 2. As can be expected by these profiles, the mean mathematics performance of Cluster 1 is much lower than the mean math performance of Cluster 2 (see Table 1).

Table 1: Characteristics of global/first level clusters

Clus- ter	**population** **size** (♀ in %)	**math score** ∅	 ♀	 ♂
1	13399687 (52%)	445	442	449
2	11321033 (48%)	468	461	475
all	24720720 (50%)	456	451	461

When we consider the relative number of students from different countries, we see that every country has students in both clusters. In fact, the distribution of the 15-year-old student population between the two clusters is quite equal in each country. For Cluster 1, the mean percentage of students from a country is 55% while for Cluster 2, the mean is 45%, and both have the standard deviation of 10. In all of the in PISA participating countries and territories, there are higher and lower performing students and it seems that they share the same characteristics. Additionally, the distribution between girls and boys is quite equal, although somewhat in favor of boys: Only 48% of the students in the cluster with the scale indices that are associated with high performance in mathematics are girls. Moreover, the average math score of the boys is in both clusters higher than the average math score of the girls (see Table 1).

3.2 Second Level

Following the approach as described above, we run the clustering algorithm again, but this time for each of the two global clusters obtained in the first level separately. According to the same rule given in Sec. 2.1, i.e., stop enlarging k during the search when all the cluster indices are increasing, we get for both of the global clusters $K = 4$ as a number for their subclusters.

3.2.1 *Subclusters of Cluster 1*

Table 2: Characteristics of subclusters of Cluster 1

Clus- ter	**population** **size** (♀ in %)	**math score** ∅	 ♀	 ♂
1-1	2792046 (56%)	439	438	440
1-2	3873035 (52%)	391	388	394
1-3	3072064 (58%)	466	464	468
1-4	3662542 (45%)	491	489	492

The subclusters of the global Cluster 1 are visualized in Fig. 3 and characterized in Table 2. If we set the threshold

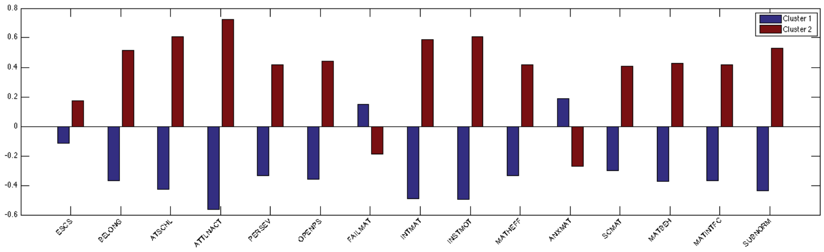


Figure 2: Characterization of the two global clusters.

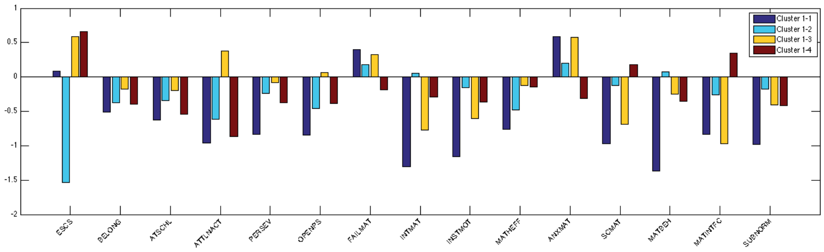


Figure 3: Characterization of the four subclusters of Cluster 1.

of how many students should at least be from one country to 21%, we obtain the following countries for the subclusters: Cluster 1-1 (i.e., subcluster 1 of Cluster 1) contains at most students from East Asia with the exception of China: More than 30% of Japan's 15-year-old student population belongs to this cluster, 26% of Korea's and and 25% of Taiwan's. The remaining students represent a mixture from many different countries which, however, are only represented by less 21% of their 15-year-old student population.

Cluster 1-2 contains almost entirely students from developing countries. Hereby, students from Vietnam form with 49% the majority. Moreover, Indonesia, Thailand (both > 30%) and Brazil, Colombia, Peru, Tunisia, and Turkey (all > 25%) are represented by this cluster. The cluster is, as can be seen from Fig. 3, most notably characterized by a very low economic, social and cultural status (*escs*). That means that the students in this cluster - as a subset of the global Cluster 1 which already represented the more disadvantaged students (see Fig. 2) - are the most disadvantaged.

Cluster 1-3 consists in the majority of students from Eastern Europe: Serbia, Montenegro, Hungary, Slovak Republic (all > 23%) and Romania (almost 22%) constitute the majority. As we can see from Fig. 3, this cluster is the only one in the group of subclusters of the global Cluster 1, that generally was characterized by negative attitudes and perceptions (see Fig. 2), which actually can be distinguished by positive attitudes towards school (*attlnact*). Moreover, it is the cluster with mainly girls in it.

Cluster 1-4 accommodates mainly students from Western and Central Europe. Most of the 15-year-old student population from the Netherlands (39%) are in this cluster, followed by Belgium with 29%, and the Czech Republic with 27%. This cluster is characterized by the highest *escs* among the students of the global Cluster 1. Furthermore, although they have negative values in most of the scale indices, they have a higher self-concept in math, and also much higher intentions to use mathematics later in life in comparison with their peers.

3.2.2 *Subclusters of global Cluster 2*

Table 3: Characteristics of subclusters of Cluster 2

Clus-	**population**	**math score**		
ter	**size** (♀ in %)	∅	♀	♂
2-1	3127958 (43%)	526	523	528
2-2	2739481 (54%)	457	457	458
2-3	3521092 (50%)	400	397	403
2-4	1932502 (44%)	515	506	523

The subclusters of the global Cluster 2 are characterized in Fig. 4 and summarized in Table 3. Again, we search for clusters that mostly deviate from the others. Cluster 2-1 is such a cluster: The students in this cluster have the highest average math score (see Table 3), the highest intentions to pursue a mathematics related career but a sense of belonging to school (*belong*) and subjective norms in mathematics (*subnorm*) that are only about the same as the average of the whole 15-year-old student population (see Fig. 4). The subjective norms in mathematics measure how people important to the students, such as their friends and parents, view mathematics. In the global Cluster 2, those students

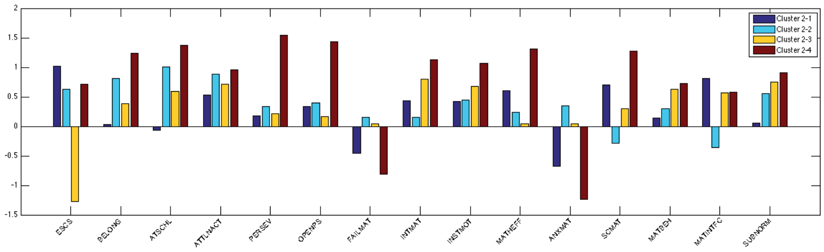


Figure 4: Characterization of subclusters of Cluster 2.

who had high positive values in the other scale indices associated with high performance in mathematics, also thought that their friends and family view mathematics as important (their *subnorm* value is very high, see Fig. 2). Students in this cluster, however, seem not to be influenced or affected by what people close to them think. It appears to be a rather strong cluster that also has the highest percentage of boys in it. For this cluster, we again compute the relative number of students from each country. And indeed, it shows a very clear country-profile. The highest percentage of students come from the English-speaking and Nordic countries: Denmark (more than 30%), Iceland and Sweden (both > 26%) have the highest percentages of their 15-year-old student population in this cluster. Followed by the two highest performing districts in the USA, namely Connecticut and Massachusetts, with both more than 25%. Besides these countries and territories, the cluster has also a high share of students from Norway, Finland, Great Britain, Australia, and Canada (almost 22% or more). Additionally, the USA has with more than 21% still a relatively high share of students in this cluster. According to the Hofstede Model (see Sec. 1), all of these countries are characterized by high individualism.

Also Cluster 2-3 shows an explicit country profile: 36% of the 15-year-old student population from India are in this cluster. Moreover, the cluster consists of students from Peru and Thailand (both 30%), Turkey (27%) and Vietnam (26%). Altogether, we find here the most disadvantaged students (indicated by the very negative *escs*) among the subgroups of the global Cluster 2 and the largest share of students come from the developing countries. However, these students have very positive attitudes towards education and show relatively high values in all scale indices that are associated with high performance in mathematics.

To this end, Cluster 2-2 and Cluster 2-4 have less obvious country affiliations. Cluster 2-2 can at best be described as containing mostly countries with Islamic culture. Most of the students are from the United Arab Emirates and Albania (both 21%), Kazakhstan and Jordan (both 19%). According to the Hofstede Model, these countries are similar in that way that they all show very high power distance. Cluster 2-4 has with 25% the highest share of students also from Kazakhstan, but the remaining countries in this cluster (all have less than 17% of their 15-year-old students population in it) are widely mixed.

Altogether, among the clusters at the second level, Cluster 2-1 appears to be the most interesting one, i.e., the most distinct group with the clearest country profiles.

3.3 Third Level

Recursively, we repeat the same approach on the next level, i.e., for the subclusters of the eight clusters identified in Sec. 3.2. For all the new subclusters, the best number of clusters as determined by the cluster indices are as follows: 6, 12, 10, and 6 for the four subclusters of the first global cluster, and 2, 8, 3, and 6 for the four subclusters of the second global cluster. This means that we have 53 different clusters on this level - almost as many as different countries/territories in the whole PISA 2012 data. If our hypothesis is true, we should be able to find clusters that clearly contain more students from certain countries. Exactly as in Sec. 3.2, we first of all compute the basic facts of each cluster. Note, however, that the deeper we go in the hierarchy the more clusters we encounter and the more difficult it becomes to define clear rules and thresholds to distinguish significant characterizations of clusters.

3.3.1 *Subclusters of Cluster 1-3*

Table 4: Characteristics of subclusters of Cluster 1-3

Cluster	**population**	**math score**		
	size (♀ in %)	∅	♀	♂
1-3-1	335240 (61%)	493	492	495
1-3-2	262779 (48%)	539	540	538
1-3-3	368591 (51%)	461	460	462
1-3-4	273629 (66%)	492	491	492
1-3-5	359721(56%)	427	428	426
1-3-6	275513 (63%)	437	436	438
1-3-7	264017 (63%)	443	441	447
1-3-8	318607 (63%)	460	457	464
1-3-9	216704 (60%)	421	418	424
1-3-10	397263 (56%)	481	482	480

The first interesting cluster appears in the 1-3 group. Cluster 1-3-8 accommodates mainly students from South West Europe: Austria, Liechtenstein, Spain, France, and Italy. According to the Hofstede Model, all of these countries are depicted by high avoidance of uncertainty.

3.3.2 *Subclusters of Cluster 1-4*

The characterization of the subclusters in the 1-4 group are provided in Fig. 6, and summarized in Table 5. Also here,

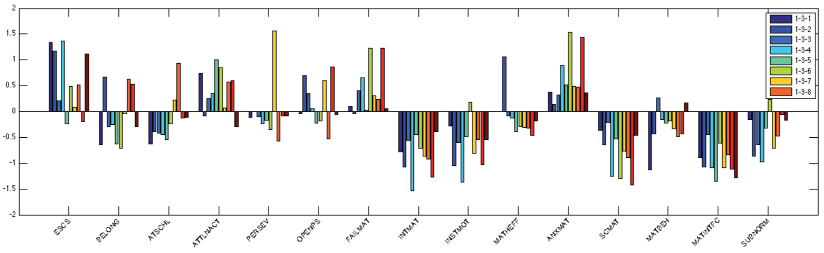


Figure 5: Characterization of subclusters of Cluster 1-3.

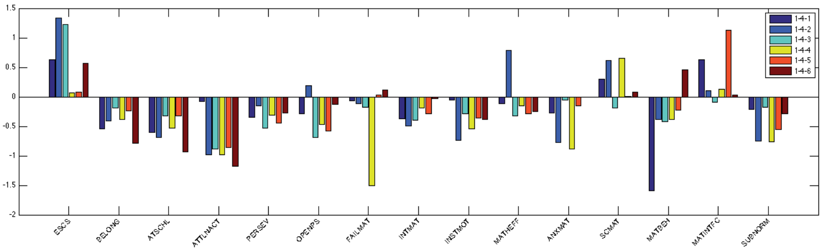


Figure 6: Characterization of the subclusters of Cluster 1-4.

Table 5: Characteristics of subclusters of Cluster 1-4

Clus- ter	population size (♀ in %)	math score ∅	 ♀	 ♂
1-4-1	485599 (48%)	481	480	482
1-4-2	520763 (38%)	556	558	555
1-4-3	771799 (53%)	494	494	495
1-4-4	489528 (43%)	497	491	501
1-4-5	754515 (48%)	470	467	473
1-4-6	640338 (38%)	461	465	458’

we are searching for explicit country clusters. This search is realized by looking at the histograms and identifying those clusters that for some countries have a considerably higher share of their 15-year-old student population in it than for the remaining countries. The histogram in Fig. 7 shows one example of this for Cluster 1-4-2: In this cluster, the portion of students in it deviates significantly from the others for exactly one country with 10% of its 15-year-old student population. This country is the Netherlands. For all other countries, the share of their 15-year-old student population in this cluster is less than 6% (see Fig. 7). As can be seen from Fig. 6, this ‘Netherlands Cluster’ is characterized by having the highest math self-efficacy amongst its group.

Cluster 1-4-1 is again a mixture of Nordic and English-speaking countries. The highest share of students in this cluster come from the United Kingdom, Ireland, Norway, New Zealand, and Sweden. As these two country profiles were already detected to be in the same cluster on the higher cluster level (see Sec. 3.2.1), it really seems that students from these countries share many similar characteristics.

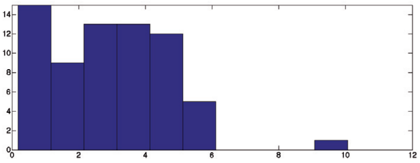


Figure 7: Histogram of the distribution of countries from the students in Cluster 1-4-2.

Cluster 1-4-4 has the highest share of East Asian countries including two of the three districts of China that participated in PISA 2012. Most of the students in this cluster come from Japan, followed by Taiwan, Macao-China and Hong Kong-China. One of the most distinct feature of this cluster is, as can be seen from Fig. 6, the high self-concept in mathematics (*scmat*). According to the Hofstede Model (see Sec. 1), all of these countries show high pragmatism.

3.3.3 *Subclusters of Cluster 2-1*

Table 6: Characteristics of subclusters of Cluster 2-1

Clus- ter	population size (♀ in %)	math score ∅	 ♀	 ♂
2-1-1	1346930 (40%)	562	557	566
2-1-2	1781028 (45%)	498	500	497

From Sec. 3.2, we concluded that Cluster 2-1 was the most interesting one. Moreover, Cluster 2-1 was the cluster that had the highest share of two country profiles in it: On the one hand, the English-speaking countries, and, on the other

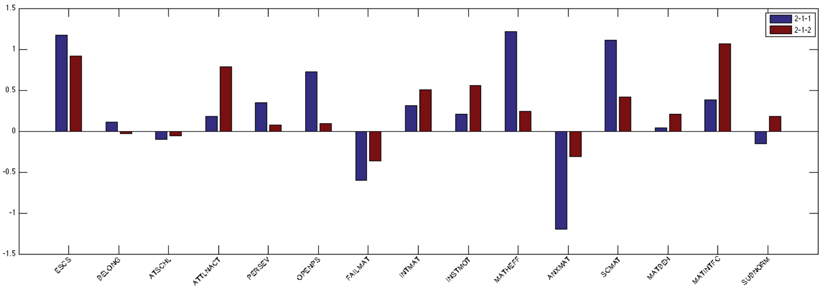


Figure 8: Characterization of subclusters of Cluster 2-1.

hand, the Nordic countries. Interestingly, the cluster indices also suggest to divide this cluster into two further countries. However, when we look again at those countries that have the highest percentages of their 15-year-old students, the two clusters still contain mostly students from both country profiles. For example, 15% of the Danish 15-year-old student population are in Cluster 2-1-1, and 14% are in Cluster 2-1-2. Similarly, 14% of the 15-year-old student population from Connecticut are in Cluster 2-1-1, and 11% in Cluster 2-1-2. Apparently, this cluster does not divide any further between Nordic and English-speaking countries. It only divides the high-performing students from these countries into two types: On the one hand, the type that has a very high self-efficacy (*matheff*) as well as self-concept (*scmat*) in math, i.e., the students that have a very high belief in their own ability, and, on the other hand, the type that has very high intentions to pursue a math related career (*matintfc*).

However, also a new clear group of countries appears. Cluster 2-1-1 has a very high share of German-speaking countries in it: More than 12% of Germany's and Switzerland's 15-year-old student population, and 10% of Austria's can be found in this cluster. None of these countries appear in the sibling Cluster 2-1-2 when the threshold is set to 9%. It seems that high-performing German-speaking students feel very confident in solving mathematical tasks but only show a moderate positive value in the intentions to use mathematics later in life, a characteristic that one would associate the most with the traditional functional German stereotype (see Sec. 1) that is expected to attach great importance to utilitarianism [4]. According to the Hofstede Model, all of these three German-speaking countries are considered to be highly masculine.

3.3.4 Subclusters of Cluster 2-4

Table 7: Characteristics of subclusters of Cluster 2-4

Clus-ter	population size (♀ in %)	math score ∅	♀	♂
2-4-1	186107 (37%)	533	528	536
2-4-2	430729 (40%)	582	575	588
2-4-3	261838 (45%)	440	436	443
2-4-4	378120 (50%)	477	468	486
2-4-5	430105 (47%)	520	519	521
2-4-6	245603 (40%)	516	500	526

The subclusters of Cluster 2-4 are summarized in Table 7 and characterized in Fig. 9. The clearest country profile among this group is 2-4-6: It consists to the highest share of students from high-performing Asian countries: Shanghai-China and Singapore. As we can see from Fig. 9, similarly to Cluster 1-4-4 (see Sec. 3.3.2) that also contained a high share of East Asian students, this cluster is characterized as well by a high self-concept in mathematics (*scmat*). The students in this cluster believe that mathematics is one of their best subjects, and that they understand even the most difficult work. Furthermore, as already found for Cluster 1-4-4, also for this cluster the main countries show high pragmatism according to the Hofstede Model.

4. CONCLUSIONS

In this article, we have introduced a clustering approach that has both partitional and hierarchical components in it. Moreover, the algorithm takes weights, aligning a sample with its population into account and is suitable for large data sets in which many missing values are present.

The hypothesis in our study was that the different clusters determined by the algorithm, when all students with their attitudes and behaviors towards education are given as input, could be explained by the country of the students in particular clusters. Our overall results on the first level showed that in each cluster students from all countries exist and that the actual test performance (as well as a simple division in positive and negative attitudes towards education) explain the clusters much better than the country from which the students in the particular cluster come from.

However, on the next two levels many clusters were detected that obviously had a much higher share of students from certain countries. For example, an Eastern Europe, a German-speaking, an East Asia, and a developing countries cluster were identified. On the second level, also a very clear cluster that consisted to a high portion of Nordic and English-speaking countries appeared. This cluster did not split further on the next level to fully separate these two distinct country profiles. Instead, the cluster was divided into two student types, of which both the Nordic as well as the English-speaking countries seem to have an almost equal share of their students from.

Summing up, we conclude that groups of similar countries,

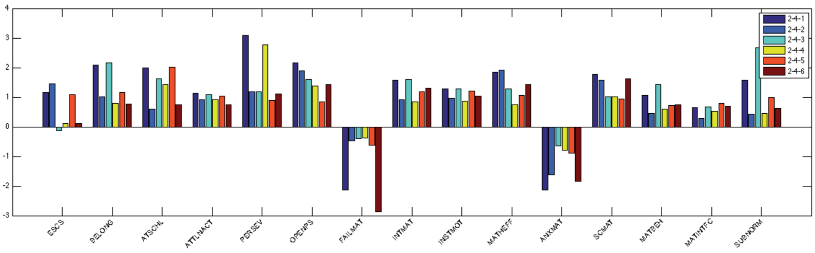


Figure 9: Characterization of subclusters of Cluster 2-4.

e.g., by means of geographical location, culture, stage of development, and dimensions according to the Hofstede Model, can be found by clustering PISA scale indices but the actual country stereotypes exist only to a very marginal extent. However, in a further work the rules how to find relevant clusters could be improved and more variables than the 15 scale indices utilized here could be included to the algorithm. The PISA scale indices are linked to math performance and in every country there are higher and lower performing students who share similar overall characteristics. Nevertheless, we think that the overall results presented here show a very promising behavior already, and we expect that the resulting clusters of our algorithm could be explained even clearer by the country of the students if additional information such as the students' temperament would be available for the clustering algorithm.

References

[1] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[2] D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[3] S. Harrell. Why do the Chinese work so hard? Reflections on an entrepreneurial ethic. *Modern China*, pages 203–226, 1985.

[4] M. F. Herz and A. Diamantopoulos. Activation of country stereotypes: automaticity, consonance, and impact. *Journal of the Academy of Marketing Science*, 41(4):400–417, 2013.

[5] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Edward Arnold, London, 1998.

[6] G. Hofstede. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture*, 2(1):8, 2011.

[7] T. Kärkkäinen. On cross-validation for MLP model evaluation. In *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science (8621), pages 291–300. Springer-Verlag, 2014.

[8] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[9] T. Kärkkäinen and M. Saarela. Robust principal component analysis of data with missing values. *To appear in the Proceedings of the 11th International Conference on Machine Learning and Data Mining MLDM*, 2015.

[10] T. Kärkkäinen and J. Toivanen. Building blocks for odd-even multigrid with applications to reduced systems. *Journal of Computational and Applied Mathematics*, 131:15–33, 2001.

[11] M. Kim and R. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[12] J. Moreno-Torres, J. Sáez, and F. Herrera. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.

[13] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[14] M. Saarela and T. Kärkkäinen. Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[15] M. Saarela and T. Kärkkäinen. Analysing Student Performance using Sparse Data of Core Bachelor Courses. *JEDM-Journal of Educational Data Mining*, 7(1):3–32, 2015.

[16] M. Saarela and T. Kärkkäinen. Weighted clustering of sparse educational data. *To appear in 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[17] P. Wartiainen and T. Kärkkäinen. Hierarchical, prototype-based clustering of multiple time series with missing values. *To appear in 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

PVI

FEATURE RANKING OF LARGE, ROBUST, AND WEIGHTED CLUSTERING RESULT

by

Mirka Saarela, Joonas Hämäläinen, Tommi Kärkkäinen 2017

Proc. of the 21th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 96–109

Feature Ranking of Large, Robust, and Weighted Clustering Result

Mirka Saarela, Joonas Hämäläinen, and Tommi Kärkkäinen
mirka.saarela,joonas.k.hamalainen,tommi.karkkainen@jyu.fi

Department of Mathematical Information Technology, P.O. Box 35 (Agora), FI-40014
University of Jyväskylä, Finland

Abstract. A clustering result needs to be interpreted and evaluated for knowledge discovery. When clustered data represents a sample from a population with known sample-to-population alignment weights, both the clustering and the evaluation techniques need to take this into account. The purpose of this article is to advance the automatic knowledge discovery from a robust clustering result on the population level. For this purpose, we derive a novel ranking method by generalizing the computation of the Kruskal-Wallis H test statistic from sample to population level with two different approaches. Application of these enlargements to both the input variables used in clustering and to metadata provides automatic determination of variable ranking that can be used to explain and distinguish the groups of population. The ranking method is illustrated with an open data and then, applied to advance the educational knowledge discovery from large scale international student assessment data, whose robust clustering into disjoint groups on three different levels of abstraction was performed in [19].

Keywords: Population analysis; Kruskal-Wallis test; Robust Clustering; Educational Knowledge Discovery

1 Introduction

Various large-scale educational assessments, like the Programme for International Student Assessment (PISA), regularly collect large amount of data characterizing worldwide student populations to assess and compare arrangements and policies between different educational systems [16]. Although data originating from these assessments are of high quality and publicly available, there is surprisingly little research activity on the secondary analysis. This is due to the technical complexities within the different representations and transformations of data and the lack of methods that allow advanced analysis of these large datasets [18]. One example of the complication of analyzing PISA datasets are the weights. Through complex sampling designs only certain students of the studied population are selected for the assessment and weights are used to indicate the number of students in the population that a sampled student represents. This means that these weights must be taken into account in all steps of the knowledge discovery to analyze the population instead of the collected sample (e.g., [20, 14]).

The purpose of this paper is to advance the educational knowledge discovery from a robust, weighted clustering result. There exists various clustering methods and approaches, like e.g. density-based, probabilistic, grid-based, and spectral clustering [2], together with their comparisons and evaluations (e.g., [6]). Although hierarchical methods allow summarization and exploration of a given dataset through the visual dendrogram, the basic form of the technique is not scalable to large number of observations because of the pairwise distance matrix requirement [25]. Moreover, it is not clear how to take into account the weights in hierarchical clustering as presented, e.g., in PISA datasets. On the other hand, in [3] a robust (cf. [24]) prototype-based clustering algorithm was developed that can handle large datasets with high and unknown sparsity patterns (i.e., tens of percents of missing values). This paper continues the efforts of [19], where the weighted enlargement of the above-mentioned algorithm was applied to create prototypes for the PISA 2012 dataset on three different levels of abstraction, with different numbers of clusters of the student population. The dynamic numbers of clusters were based on the use of multiple cluster indices (e.g., [13]) suggesting the number of clusters, again taking into account the weights (see [19] for details).

One main advantage of crisp, prototype-based clustering result is the guarantee of globally separable subsets of data. The data division is completely determined by the disjoint labels, typically integers from 1 to K for K clusters, encoding the clustering result. This means that, in order to make an interpretation of the result, one can consider and compare data distributions of both the actual variables used in clustering as well as relevant metadata. Note that the use of a hierarchical clustering method with locally greedy aggregation could produce clusters of arbitrary shape in the data space, which could then be difficult or even impossible to interpret because of the overlapping variable distributions.

The results in [19] were obtained with a robust clustering method with (available data) spatial median as the cluster prototype, which is characterized by the Laplace density distribution. A feature selection approach for the robust EM-algorithm with Laplace mixture models was suggested in [5]. There the feature selection, similarly to the construction of classifiers [11], referred to ranking the given input features to select the most important ones for the clustering result. Here, our purpose is, similarly to the techniques proposed in [23, 4], to assess the importance of variables with a given labeling. For this purpose, we apply the same method as in [5] where it was suggested that the feature ranking can be realized by Kruskal-Wallis (KW) statistical test. More precisely, the estimate of importance of a random variable with clustering provided labeling is supplied by the H statistics of the KW test [15], without need to compute the p-values and perform the actual statistical testing. To omit the hypothesis testing relaxes both the requirements of the KW test concerning the equal variances [15] and selection of appropriate distribution for the test statistics [21]. Moreover, because KW is a univariate method, it is easy to restrict the computation of the test statistic to the available values of a variable. This means utilizability with an arbitrary sparsity pattern.

Hence, one needs to generalize the KW H into the population level by using the weights. This is a difficult problem in statistics because of the reliance of KW on data ranking. After an extensive search for relevant literature and knowledge we were able to identify one related work generalizing KW [1], but not solving the problem at hand.

The only article that was identified as fully relevant was [22], which suggested a very natural generalization of KW for *integer weights*: create univariate data to compute the KW test statistic, where each observation is copied as many times as the integer weight suggests. Clearly, we then precisely test the target population and not the sample. The purpose of this paper is to propose an approximate extension of this approach to real-valued weights, by utilizing the classical bootstrapping [8], and to compare this to an analytically derived novel heuristic formula. Both of these approaches are tested and evaluated with two different existing clustering results from [19], when ranking both actual input variables and selected set of metadata variables.

2 On PISA data

The collected data of each PISA assessment, which since 2000 is conducted every three years, can be downloaded from the website[1] of the Organisation of Economical and Cultural Development (OECD). To select a reliable sample of the population, which in PISA are all 15-year-old students within the participating countries, the OECD applies a two-stage sampling design: First, schools attended by 15-year-old students are assigned to mutually exclusive groups based on explicit strata and schools from these groups are selected with probabilities proportional to their size. Then, students within those school are selected randomly with equal probability. The weight w_i assigned to each participating student i consists of the school base weight, the within-school base weight, and five adjustment factors, especially the one which compensates the non-participation of a sampled student [17]. Students that are sampled for the PISA test are asked to show their proficiencies in a cognitive test and answer a background questionnaire, which gathers information about demographics, activities, and attitudes of the students.

Table 1 details all PISA 2012 variables used in this study. The left-hand side of the table shows all the variables that in [19] were clustered on a population-level. The *ESCS* combines all information of the PISA background questionnaire that relate to the students' economic, social and cultural situation. The next five variables on the left-hand side of Table 1 are generally associated with the students' success in the PISA cognitive test, and the remaining nine variables relate directly to the students' mathematics performance, which was the main assessment area in PISA 2012. All of these 15 variables are so-called PISA scale indices that summarize many of the original questions in the students' background questionnaires by employing the Rasch model [17]. Since only a subset of all test item are allocated to each student (this is called rotated design), around one third of the values for these 15 variables are missing.

On the right-hand side of Table 1, the meta-variables to be used in this study are listed. The first eight variables of general interest are all PISA scale indices that were computed to summarize the information obtained from the ICT questionnaire, which assessed the students' computing availability and familiarity as well as their attitudes towards computers. The next and last set of variables in Table 1 are the plausible values (PVs) for each assessment domain (mathematics, reading, and science). PISA does not provide individual test performance scores. Instead, to reliably assess the proficiencies

[1] https://www.oecd.org/pisa/pisaproducts/

of populations, five PVs for each assessment domain are estimated with Bayesian statistics and reported for each student. Note that we have allocated only one line in the table per assessment domain for the three sets of PVs but there are five single PVs vectors per assessment domain, i.e., 15 PVs altogether, that are used in the analysis.

Table 1. PISA variables used in this study with the original variables (i.e., the data that was used for clustering) on the left-hand side and metadata (i.e., additional PISA variables used to explain the clustering result) on the right-hand side.

PISA data used for clustering		PISA metadata	
variable	**ID**	**variable**	**ID**
economic, social and cultural status	ESCS	ICT availability at home	ICTHOME
sense of belonging	BELONG	ICT availability at school	ICTSCH
attitude towards school: learning outcome	ATSCHL	ICT entertainment use	ENTUSE
attitude towards school: learning activities	ATTLNACT	ICT use at home for school-related tasks	HOMSCH
perseverance	PERSEV	use of ICT at school	USESCH
openness to problem solving	OPENPS	use of ICT in math lessons	USEMATH
self-responsibility for failing in math	FAILMAT	positive attitudes towards computers	ICTATTPOS
interest in mathematics	INTMAT	positive attitudes towards computers	ICTATTPOS
instrumental motivation to learn math	INSTMOT	plausible values 1-5 in mathematics	PVMATH
self-efficacy in mathematics	MATHEFF	plausible values 1-5 in reading	PVREADING
anxiety towards mathematics	ANXMAT	plausible values 1-5 in science	PVSCIENCE
self-concept in math	SCMAT		
behaviour in math	MATBEH		
intentions to use math	MATINTFC		
subjective norms in math	SUBNORM		

The PVs are random draws from the Bayesian posterior distribution of a student's ability. In PISA, the prior distribution is a population model that is estimated with a latent regression model. This latent regression computes the average proficiencies of examinee subgroups given evidence about the distribution and associations of collateral variables in the data. In PISA 2012, these collateral variables included to the latent regression model were all available student-level information besides their performance in the cognitive test [17, page 157]. That means, in particular, that also all variables listed in Table 1 except the 15 PVs themselves have been used to estimate the PVs, and therefore, the PVs cannot be seen totally independent of them. The likelihood of the success in test is a Rasch model, where the probability of success is a logistic function of the latent ability and some parameters (e.g. difficulties) of the test items. The obtained posterior distribution of a student's ability is specific for each student, since each student has different values of background variables and test results.

To sum up, student proficiencies in PISA are not directly observed. The PVs are estimates for group performance and only a selection of likely proficiencies for students that attained each score. Moreover, for the study at hand, it is important to note that all background information (i.e., all data that were clustered and all metadata except the PVs themselves) have been used in the latent regression model which contributes to the posterior distribution from which the PVs are drawn from.

3 Methods and formulations

Let $\{x_i\}_{i=1}^N$ be a given, multidimensional dataset, where N observations $x_i \in \mathbb{R}^n$ are given. Assume further that a given set of positive, real-valued weights $\{w_i\}_{i=1}^N$ is also given. Moreover, assume that there is a set of missing values in $\{x_i\}$ with unknown sparsity pattern. To identify this pattern, define the projection vectors $p_i, i = 1, \ldots, N$, that capture the existing variable values:

$$(p_i)_j = \begin{cases} 1, \text{if } (x_i)_j \text{ exists}, \\ 0, \text{otherwise}. \end{cases} \tag{1}$$

3.1 Robust, prototype-based clustering method for weighted sparse data

Let us briefly recapitulate the clustering method and the overall approach that was used hierarchically in [19], to produce three levels of disjoint clusters of PISA 2012 population with 2, 8, and 53 clusters, respectively.

The spatial median clustering algorithm, `k-SpatMeds`, proceeds similarly to any prototype-based method: first, an initial set of *complete* (i.e., no missing values) prototypes is created and second, these are refined by iteratively linking observations to the closest prototype whose value is then recomputed. The algorithm stops when there are no more changes in the linking. Mathematically, the score function that is locally minimized via the search procedure reads as follows:

$$\mathcal{J}_w = \sum_{j=1}^{K} \sum_{i=1}^{n_j} w_i \|\mathrm{Diag}\{p_i\}(x_i - c_j)\|_2. \tag{2}$$

Here, Diag transforms a vector into a diagonal matrix. The latter sum is computed over the subset of data attached to the jth cluster. One observes from (2) that to take into account the first-order alignment of the sample data with the corresponding population is straightforward. Moreover, projection of the Euclidean distance between the observation and the prototype to available values creates an implicit (secondary) weighting that favors more complete observations over the sparser ones in cluster creation. Algorithmically, one still needs to check that the iterative refinement of the prototypes does not introduce missing values to them, because the resulting set of cluster prototypes $\{c_i\}_{i=1}^K$ should be complete to allow proper interpretation. The robustness of this algorithm as thoroughly described and tested in [3], refers to the tolerance of both missing values and noisy data. To this end, one can apply the `k-SpatMeds` algorithm hierarchically to refine a set of disjoint clusters further.

3.2 Construction of test statistic for Kruskal-Wallis with weights

Next we describe two different approaches to estimate the test statistic H of the KW rank-test with real-valued weights. Because the KW test is univariate, we can restrict ourselves to univariate random variable.

Integer approximation with bootstrapping Let $\{x_i, l_i\}_{i=1}^N$ be the pairs of a univariate observation $x_i \in \mathbb{R}$ and its cluster-indicating label $l_i \in \mathbb{N}$, where $1 \leq l_i \leq K$ for K denoting the number of clusters/groups. Let $n_k = |C_k| = \{i \in \mathbb{N} \,|\, l_i = k\}$ determine the size of cluster C_k. The original formula for the KW H is given by [15]

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{s_k^2}{n_k} - 3(N+1), \tag{3}$$

where r_i denotes the *rank* of observation x_i in global sorting and $s_k = \sum_{i \in C_k} r_i$ the sum of ranks in cluster C_k. When there are equal values (ties) in data, one can compute the mean rank of equal observations and share this value among the ties.

As described, $w_i \in \mathbb{R}$ measures the amount of population that the ith observation represents. If all w_i's are integers, then in [22] it was proposed how to modify the basic KW test: rank a derived dataset representing the whole population, where each (available) observation is copied as many times as the weight suggests. This approach is referred from now on as *Integerweighted-KW, IW-KW*. Note that when such an enlarged data are ranked we end up with multiple ties whose mean ranks are then shared. In the following, we describe a novel approach how to approximate this integer-weighted KW using a bootstrapping technique.

Let w denote an arbitrary, real-valued weight. The proposed technique is, firstly, based on approximating w up to an accuracy of the first decimal place. This can be simply done as follows: determine the two integers $w_l = \lfloor w \rfloor$ and $w_h = \lceil w \rceil$ that provide lower and upper bound of w as integers. Let then $d = [10 * (w - w_l)]$ be the rounded integer that encapsulates the decimal place 1 of w. Vector v of ten integers, which is created by repeating w_l $10 - d$ times and w_h d times, provides an integer-approximating set of real-valued w in such a way that the mean of v is exactly the same as w up to the first decimal. For instance, for $w = 8.647$, $w_l = 8, w_h = 9$, and $d = 6$. And, for $v = \begin{bmatrix} 8\ 8\ 8\ 8\ 9\ 9\ 9\ 9\ 9\ 9 \end{bmatrix}$, we have $mean\{v\} = 8.6$. Similarly, in order to create an integer-approximation of w being accurate to the second decimal place, it is enough to just redefine $d = [100 * (w - w_l)]$. Proceeding with the example just given, the integer vector of size 100 with 65 nines and 35 eights would yield to $mean\{v\} = 8.65$. For the general procedure, the result of the just proposed integer approximation of all weights is stored in the matrix $\mathbf{W} \in \mathbb{N}^{N \times D}$, where D is 10 when approximating the first decimal place and 100 for the second decimal place, correspondingly.

Next we suggest to use the classical bootstrapping [8] to create a set of KW test statistics based on the IW-KW and W. Hence, we create a random sample of indices $\{1, \ldots, N\}$ with replacement, and for the resulting unique set of indices $\tilde{I}$, for the available values of $\{x_i\}_{i \in \tilde{I}}$, we apply IW-KW. When this is repeated D times for all the integer columns of W, we obtain D different samples of the bootstrap estimate of the KW H. To this end, similarly as with the derivation of W, we then simply take the mean of the D-vector to produce the final approximation of H for the real-valued weights.

Analytic formula Let $\bar{r}$ denote the global mean rank (equal to $\frac{1+N}{2}$) and $\bar{r}_k$ the mean rank of the observations in cluster C_k. An equivalent form of the original formula (3) for the KW test statistic H, as given in [9], reads as

$$H = (N-1)\frac{\sum_{k=1}^{K} n_k(\bar{r}_k - \bar{r})^2}{\sum_{i=1}^{N}(r_i - \bar{r})^2}. \tag{4}$$

From this form, it is easy to derive an interpretation of the KW test statistic. With clusterwise $\bar{r}_k$ and global $\bar{r}$ mean ranks, the dividend presents sum of clusterwise variances multiplied by the size of the cluster whereas the divisor computes the global variance of ranks. Hence, when the weights represent the number of samples in the population, it is straightforward to derive an analogous formula to (4) in the population level. Hence, let $\bar{r}_w = \frac{\sum_{i=1}^{N} w_i r_i}{\sum_{i=1}^{N} w_i}$ be the weighted average rank and $(\bar{r}_w)_k$ the weighted average rank of cluster C_k. Then, we define

$$H_w = \frac{\sum_{k=1}^{K}(\sum_{i\in C_k} w_i)((\bar{r}_w)_k - \bar{r}_w)^2}{\sum_{i=1}^{N} w_i(r_i - \bar{r}_w)^2}. \tag{5}$$

Note that we have omitted the multiplier $(N-1)$ from (4), which would be generalized into $(\sum_i w_i - 1)$ to represent the whole population. With PISA 2012 weights, which align the half a million students sample to the 24 million population, this means we do not include multiplication of H_w by over 24 million. Because the final ranking of variables, as suggested in [5], is based on sorting the H values of the variables in descending order, this omission does not change the result.

4 Evaluation

Implementation We computed the KW rank-test H test statistics for real-value weighted data with two approaches, as described in Section 3. The bootstrapping with the IW-KW was tested with two different Ws. We will refer to the bootstrapping based method as Bootstrap KW. Further, Bootstrap KW with D = 10 refers to the one decimal place approximation of real-valued weights. Similarly, the two decimal place approximation is referred as Bootstrap KW with D = 100. In addition, the KW test statistics were computed directly from formula (5). In the following, this is shortly referred as Analytic KW. The two clustering results that are used in the experiments corresponded to 8 (*Labels 1*) and 53 (*Labels 2*) clusters from [19] in the second and third levels of refinement, respectively. The first result in [19] with the two clusters is excluded here, since the KW rank-test exactly generalizes the MannWhitney U-test for the two groups.

To speed up the computations, we implemented a parallel version of Bootstrap KW with Matlab PCT, SPMD blocks and message passing functions. The tests were run in Matlab 8.5.0 environment by using a cluster of 8 nodes. Each node consists of Intel Xeon CPU E7-8837 with 8 cores and 128 GB RAM. Each worker in the distributed computations corresponds to one of the 64 cores. Since Bootstrap KW computes the KW H values independently for each variable in a loop, those loop iterations can be easily parallelized with SPMD blocks. First, each worker reads one column of variable values from the data matrix and the corresponding sparsity indicator (1). Next, each worker computes the KW H values by utilizing its local data. Finally, results are aggregated and rankings for the variables based on the H values are formed. The number of workers is equal to the number of variables in all parallel runs.

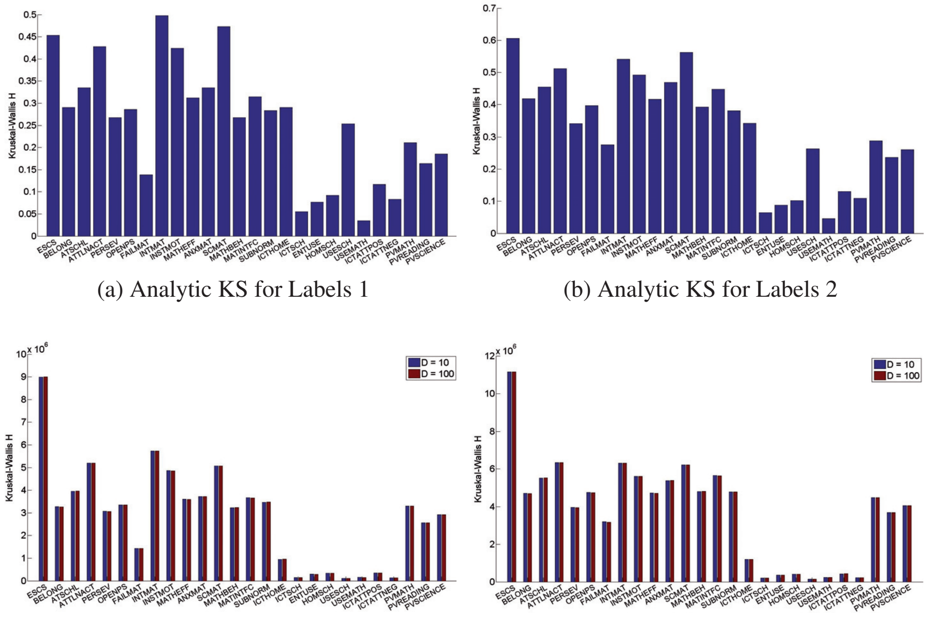
The five individual PVs for mathematics, reading, and science, as given in Table 1, were first treated as independent variables, such that five H values were computed for them. The final value of the test statistic was then taken as the mean of these according to the recommended way of analysis in [17].

Results To generally test the proposed approaches, we first used the Iris data from UCI machine-learning repository. For this, we created random integer weights in the range 5–25 and newly generated the data for each run. The KW H values for Analytic KW and Bootstrap KW $D = 100$ approaches gave the same variable ranking results in eight out of ten runs. After adding 5% zero-mean uniformly distributed noise to make weights real-values, we obtained the same ranking order for the different approaches in nine out of ten runs. Moreover, similarly as in [7], features 4 and 3 were always selected as the important ones while features 1 and 2 were always last in the list. When we used the same data for each run the ranking order was always the same.

Table 2 summarizes all ranking for the combined (originally clustered and meta) PISA data. In the table, the last column *rank of rankings* indicates for each variable the total rank, i.e. the rank of the sum of rankings of all methods on both labeling levels.

Table 2. Rankings for full (original and metadata) variables for the different analysis approaches for both PISA clustering results.

	Labels 1			Labels 2			
		Bootstrap KW			Bootstrap KW		rank of
Variable	Analytic KW	$D = 10$	$D = 100$	Analytic KW	$D = 10$	$D = 100$	rankings
ESCS	3	1	1	1	1	1	1
BELONG	11	13	13	9	13	13	12
ATSCHL	7	6	6	7	7	7	6
ATTLNACT	4	3	3	4	2	2	3
PERSEV	15	15	15	15	16	16	15
OPENPS	12	11	11	11	11	11	11
FAILMAT	20	18	18	17	18	18	19
INTMAT	1	2	2	3	3	3	2
INSTMOT	5	5	5	5	6	6	5
MATHEFF	9	9	9	10	12	12	9
ANXMAT	6	7	7	6	8	8	7
SCMAT	2	4	4	2	4	4	4
MATHBEH	14	14	14	12	9	9	13
MATINTFC	8	8	8	8	5	5	8
SUBNORM	13	10	10	13	10	10	10
ICTHOME	10	19	19	14	19	19	17
ICTSCH	25	24	24	25	25	25	25
ENTUSE	24	22	22	24	22	22	22
HOMSCH	22	21	21	23	21	21	21
USESCH	16	26	26	18	26	26	23
USEMATH	26	23	23	26	23	23	24
ICTATTPOS	21	20	20	21	20	20	20
ICTATTNEG	23	25	25	22	24	24	26
PVMATH	17	12	12	16	14	14	14
PVREADING	19	17	17	20	17	17	18
PVSCIENCE	18	16	16	19	15	15	16



(a) Analytic KS for Labels 1

(b) Analytic KS for Labels 2

(c) Bootstrap KS for Labels 1

(d) Bootstrap KS for Labels 2

Fig. 1. KW H values for two clustering results for the combined (originally clustered and meta) PISA data determined with the analytic and the two bootstrap KW approaches.

KW H values for both clustering results are shown in Figure 1. As can be seen from Table 2, variable rankings between the analytic and the bootstrapped results are highly similar with the exception that variable *USESCH* had a ranking difference 10 for Labels 1 and ranking difference 8 for Labels 2. In addition, variable *ICTHOME* had ranking difference 9 for Labels 1 and ranking difference 5 for Labels 2.

The Kendall's tau distance (see [10]) provides a way to compute distance between two ranking lists with an equal set of variables. The Kendall's tau distance is equal to the bubble sort algorithm steps to convert one list to the same order as the other one. If m is the number of elements in the list, then the maximum value for the Kendall's tau distance is $m(m-1)/2$ which is typically used to normalize this distance metric. Thus, the Kendall's tau distance is limited to an interval $[0,1]$, where value 0 refers to the identical lists and value 1 to the case where one list is the reverse of the other list. The Kendall's tau distances between the Analytic KW and Bootstrap KW with D = 100 were 0.1015 for Labels 1 and 0.1138 for Labels 2. This concludes that, overall, the rankings are highly similar as measured by the Kendall's tau distance.

Bootstrap KW with D = 10 and Bootstrap KW with D = 100 gave identical rankings for the variables. Experimentally, it seems that approximation of the real-valued weights using just the first decimal place ($D = 10$) is accurate enough. However, for a few variables slight differences can be noticed from the Figures 1c and 1d. We also computed speedups for the distributed Bootstrap KW. We measured running time for the first variable computations by using a serial implementation of the Bootstrap KW, and multiplied this with the total number of variables to get an estimate for the serial implementation running time. Further, we measured running time for the corresponding parallel implementation. Thus, parallel Bootstrap KW with $D = 100$ gives 34 × speedup compared to sequential code for Labels 1 and 35 × speedup for Labels 2. Correspondingly, parallel Bootstrap KW with $D = 10$ gives 28 × speedup for Labels 1 and 33 × speedup for Labels 2. In practice, this means that using the distributed version enables one to carry out the whole cluster analysis chain in realtime.

As expected, we see from Table 2 and Figure 1 that the actually clustered variables generally contribute more to the clustering result than the metadata variables. However, this first observation does not hold for all variables: The metadata PVs in mathematics were more important than the level of self-responsibility for failing in mathematics (see row *FAILMAT* in Table 2), which was clustered. Generally, the PVs are the most important variables from the metavariables. This ranking result makes sense because the clustered variables are, as explained in Section 2, part of the posterior model from which the PVs were sampled. Moreover, most of the clustered variables are directly associated with the students' mathematics proficiencies. Hence, the PVs in mathematics should be important variables when explaining the clustering result and, thus, these observations support the validity of our results.

As can be seen in Table 2, the students' *ESCS* is the most important variable determining the different clusters. This was already assumed in [19] where the most distinguishing country clusters were those that showed different stages of development. Moreover, the students' *ESCS* is the single variable in the whole PISA data, which accounts for most of the variance in performance [16]. Therefore, it is reasonable to assume that the variable that explains the mathematics proficiency the most, is also the most important when variables associated with the mathematics performance, are clustered. The students' *ESCS* takes not only the highest parental education and occupation into account but also the students' home possessions. Therefore, the *ICTHOME*, which summarizes the home possessions in the ICT area, is partly associated with the students' *ESCS* [17, page 132]. Hence, it seems reasonable that *ICTHOME* is next to the PVs one of the most important variables from the metadata (see Table 2).

To sum up, weighted enlargements with all approaches proposed in Section 3 successfully enabled ranking of input and metadata. Triangulation for both actual input and metadata by using two clustering results of a PISA dataset and two different algorithms/formulae showed very similar results for all methodological approaches and also for the two clustering results that were analyzed. Hence, it seems that the interpretation is not an artifact of the method used to analyze the data or only a result of the particular sample, but reflects genuine and overarching aspects of the data [12].

5 Discussion and conclusions

Large scale educational assessment data provide interesting and high quality resources for educational knowledge discovery. Although the data from these assessments are made available to the public a scarce pool of research outcomes exist that make use of those rich datasets because of the technical difficulties in them. Only one study [19] was identified, in which the whole PISA 2012 contextual data were clustered by taking the complexities of these data (especially the sparsity and the weights) into account. However, the work in [19] lacked a clear frame how to assess the importance of individual variables to interpret the clustering results.

In this study, we proposed weighted enlargements of the KW H test with different approaches, which as an independent statistical problem is not trivial. All approaches successfully enabled ranking of input and metadata. In particular, when applied to the two clustering results in [19], all approaches supported the finding that the students' *ESCS* is the most important variable determining the clusters—a fact that was also hypothesized in [19] but could not be statistically shown in there. Moreover, also the ranking of the other variables seem to support the interpretations made in [19].

The y-scales of Figures 1c and 1d illustrate the very large size of the KW test statistic(s) H for a large population, which in our case is characterized by over 24 million students worldwide. Hence, even if the nonparametric KW test can be used for testing large samples [9], the actual hypothesis testing seems practically useless. We tested the computation of the p-values for the original sample, for both clustering results and for all data and metadata variables, and found in each case that the p-value was equal to zero up to six decimal places. Hence, the hypothesis test itself does not provide any useful information for educational knowledge discovery.

Based on the high similarity of the results of the different ranking approaches, we suggest the direct KW formula with weights to be used for quick evaluation of significance of a variable on the population level. If the weighted estimates are used to derive, e.g., confidence intervals for the test statistics and the resulting rankings, the bootstrap-based approach should be used. This approach is also better aligned to the existing literature [8, 5, 22]. To this end, we conclude that the proposed approach supports quantified educational knowledge discovery from PISA and similar large-scale educational datasets.

Acknowledgments

The authors would like to thank PhD Salme Kärkkäinen for her kind and valuable suggestion to use bootstrapping.

References

1. Acar, E.F., Sun, L.: A Generalized Kruskal–Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies. Biometrics 69(2), 427–435 (2013)
2. Aggarwal, C.C., Reddy, C.K.: Data clustering: algorithms and applications. CRC Press (2013)
3. Äyrämö, S.: Knowledge Mining Using Robust Clustering, Jyväskylä Studies in Computing, vol. 63. University of Jyväskylä (2006)
4. Ceccarelli, M., Maratea, A.: Assessing Clustering Reliability and Features Informativeness by Random Permutations. In: Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, XVII Italian Workshop on Neural Networks, Proceedings. pp. 878–885. Springer (2007)

5. Cord, A., Ambroise, C., Cocquerez, J.P.: Feature selection in robust clustering based on Laplace mixture. Pattern Recognition Letters 27(6), 627–635 (2006)
6. Crabtree, D., Andreae, P., Gao, X.: QC4 - A Clustering Evaluation Method. In: Advances in Knowledge Discovery and Data Mining: 11th Pacific-Asia Conference, Proceedings. pp. 59–70. Springer (2007)
7. Dash, M., Liu, H.: Feature selection for clustering. In: Advances in Knowledge Discovery and Data Mining: 4th Pacific-Asia Conference, Proceedings. pp. 110–121. Springer (2000)
8. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7, 1–26 (1979)
9. Elamir, E.A.: Kruskal-Wallis Test: A Graphical Way. International Journal of Statistics and Applications 5(3), 113–119 (2015)
10. Fagin, R., Kumar, R., Sivakumar, D.: Comparing Top K Lists. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 28–36. Society for Industrial and Applied Mathematics (2003)
11. Fung, P.C.G., Morstatter, F., Liu, H.: Feature Selection Strategy in Text Classification. In: Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, Proceedings. pp. 26–37. Springer (2011)
12. Gifi, A.: Nonlinear multivariate analysis. Wiley (1991)
13. Kim, Y., Lee, S.: A Clustering Validity Assessment Index. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 602–608. Springer (2003)
14. Koskela, A.: Exploring the differences of Finnish students in PISA 2003 and 2012 using educational data mining. Jyväskylä Studies in Computing, University of Jyväskylä (2016)
15. Kruskal, W., Wallis, W.: Use of Ranks in One-Criterion Variance Analysis. Journal of the American statistical Association 47(260), 583–621 (1952)
16. OECD: PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II). PISA, OECD Publishing (2013)
17. OECD: PISA 2012 Technical Report. OECD Publishing (2014)
18. Rutkowski, L., Rutkowski, D.: Getting It ”Better”: The Importance of Improving Background Questionnaires in International Large-Scale Assessment. Journal of Curriculum Studies 42(3), 411–430 (2010)
19. Saarela, M., Kärkkäinen, T.: Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In: Proceedings of the 8th International Conference on Educational Data Mining. pp. 156–163 (2015)
20. Saarela, M., Kärkkäinen, T.: Weighted Clustering of Sparse Educational Data. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. pp. 337–342 (2015)
21. Spurrier, J.D.: On the null distribution of the Kruskal–Wallis statistic. Nonparametric Statistics 15(6), 685–691 (2003)
22. Tölgyesi, C., Bátori, Z., Erdős, L.: Using statistical tests on relative ecological indicator values to compare vegetation units–Different approaches and weighting methods. Ecological Indicators 36, 441–446 (2014)
23. Verde, R., Lechevallier, Y., Chavent, M.: Symbolic clustering interpretation and visualization. The Electronic Journal of Symbolic Data Analysis 1(1) (2003)
24. Yang, H., Zhao, D., Cao, L., Sun, F.: A Precise and Robust Clustering Approach Using Homophilic Degrees of Graph Kernel. In: Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, Proceedings. pp. 257–270. Springer (2016)
25. Zaki, M.J., Meira Jr., W.: Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press (2014)

PVII

ROBUST PRINCIPAL COMPONENT ANALYSIS OF DATA WITH MISSING VALUES

by

Tommi Kärkkäinen, Mirka Saarela 2015

Proc. of the 11th International Conference on Machine Learning and Data
Mining in Pattern Recognition, pp. 140–154

Robust Principal Component Analysis of Data with Missing Values

Tommi Kärkkäinen and Mirka Saarela

University of Jyväskylä, Department of Mathematical Information Technology
40014, Jyväskylä, Finland

Abstract. Principal component analysis is one of the most popular machine learning and data mining techniques. Having its origins in statistics, principal component analysis is used in numerous applications. However, there seems to be not much systematic testing and assessment of principal component analysis for cases with erroneous and incomplete data. The purpose of this article is to propose multiple robust approaches for carrying out principal component analysis and, especially, to estimate the relative importances of the principal components to explain the data variability. Computational experiments are first focused on carefully designed simulated tests where the ground truth is known and can be used to assess the accuracy of the results of the different methods. In addition, a practical application and evaluation of the methods for an educational data set is given.

Keywords: PCA, Missing Data, Robust Statistics

1 Introduction

Principal component analysis (PCA) is one of the most popular methods in machine learning (ML) and data mining (DM) of statistical origin [12]. It is typically introduced in all textbooks of ML and DM areas (e.g., [1, 10]) and is used in numerous applications [15]. It seems that the versatile line of utilization has also partly redefined the original terminology from statistics: in ML&DM, the computation of principal components and their explained variability of data, many times together with dimension reduction, is referred to as PCA, even if the term *analysis*, especially historically, refers to statistical hypothesis testing [12]. However, nowadays the use of the term PCA points to the actual computational procedure. Certainly one of the appealing facets of PCA is its algorithmic simplicity with a supporting linear algebra library: a) create covariance matrix, b) compute eigenvalues and eigenvectors, c) compute data variability using eigenvalues, and, if needed, transform data to the new coordinate system determined by the eigenvectors. This is also the algorithmic skeleton underlying this work.

Even if much researched, the use of PCA for sparse data with missing values (not to be mixed with sparse PCA referring to the sparsity of the linear model [6]) seems not to be a widely addressed topic, although [27] provides a comparison of a set of second-order (classical) methods. We assume here that there is no further information on the sparsity pattern so that the non-existing subset of data is *missing completely at random* (MCAR) [18]. As argued in [24, 25], a missing value can, in principle, represent any value from the possible range of an individual variable so that it becomes difficult to justify assumptions on data or error normality, which underlie the classical PCA that is based on second-order statistics. Hence, we also consider the so-called nonparametric, robust statistical techniques [13, 11], which allow deviations from normality assumptions while still producing reliable and well-defined estimators.

The two simplest robust estimates of location are median and spatial median. The median, a middle value of the ordered univariate sample (unique only for odd number of points, see [16]), is inherently one-dimensional, and with missing data uses only the available values of an individual variable from the marginal distribution (similarly to the mean). The spatial median, on the other hand, is truly a multidimensional location estimate and utilizes the available data pattern as a whole. These estimates and their intrinsic properties are illustrated and more thoroughly discussed in [16]. The spatial median has many attractive statistical properties; particularly that its breakdown point is 0.5, that is, it can handle up to 50% of the contaminated data, which makes it very appealing for high-dimensional data with severe degradations and outliers, possibly in the form of missing values. In statistics, robust estimation of data scattering (i.e., covariability) has been advanced in many papers [19, 28, 7], but, as far as we know, sparse data have not been treated in them.

The content of this work is as follows: First, we briefly derive and define basic and robust PCA and unify their use to coincide with the geometrical interpretation. Then, we propose two modifications of the basic robust PCA for sparse data. All the proposed methods are then compared using a sequence of carefully designed test data sets. Finally, we provide one application of the most potential procedures, i.e., dimension reduction and identifying the main variables, for an educational data set, whose national subset was analyzed in [24].

2 Methods

Assume that a set of observations $\{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbf{R}^n$, is given, so that N denotes the number of observations and n the number of variables, respectively. To avoid the low-rank matrices by the form of the data, we assume that $n < N$. In the usual way, define the data matrix $\mathbf{X} \in \mathbf{R}^{N \times n}$ as $\mathbf{X} = \left(\mathbf{x}_i^T\right), i = 1, \ldots, N$.

2.1 Derivation and interpretation of the classical PCA

We first provide a compact derivation underlying classical principal component analysis along the lines of [4]. For the linear algebra, see, for example, [8]. In general, the purpose of PCA is to derive a linear transformation to reduce the dimension of a given set of vectors while still retaining their information content (in practice, their variability). Hence, the original set of vectors $\{\mathbf{x}_i\}$ is to be transferred to a set of new vectors $\{\mathbf{y}_i\}$ with $\mathbf{y}_i \in \mathbf{R}^m$, such that $m < n$ but also $\mathbf{x}_i \sim \mathbf{y}_i$ in a suitable sense. Note that every vector $\mathbf{x} \in \mathbf{R}^n$ can be represented using a set of orthonormal basis vectors $\left[\mathbf{u}_1 \ldots \mathbf{u}_n\right]$ as $\mathbf{x} = \sum_{k=1}^{n} z_k \mathbf{u}_k$, where $z_k = \mathbf{u}_k^T \mathbf{x}$. Geometrically, this rotates the original coordinate system.

Let us consider a new vector $\tilde{\mathbf{x}} = \sum_{k=1}^{m} z_k \mathbf{u}_k + \sum_{k=m+1}^{n} b_k \mathbf{u}_k$, where the last term represents the residual error $\mathbf{x} - \tilde{\mathbf{x}} = \sum_{k=m+1}^{n} (z_k - b_k)\, \mathbf{u}_k$. In case of the classical PCA, consider the minimization of the least-squares-error:

$$\mathcal{J} = \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^T \,(\mathbf{x}_i - \tilde{\mathbf{x}}_i) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=m+1}^{n} (z_{i,k} - b_k)^2. \quad (1)$$

By direct calculation, one obtains $b_k = \mathbf{u}_k^T\, \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the sample mean. Then (1) can be rewritten as $((\mathbf{u}^T\, \mathbf{v})^2 = \mathbf{u}^T\, (\mathbf{v}\, \mathbf{v}^T)\, \mathbf{u}$ for vectors $\mathbf{u}, \mathbf{v})$ so that

$$\mathcal{J} = \frac{1}{2} \sum_{k=m+1}^{n} \sum_{i=1}^{N} \left(\mathbf{u}_k^T (\mathbf{x}_i - \bar{\mathbf{x}})\right)^2 = \frac{1}{2} \sum_{k=m+1}^{n} \mathbf{u}_k^T \,\Sigma\, \mathbf{u}_k, \quad (2)$$

where Σ is the sample covariance matrix

$$\Sigma = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})\, (\mathbf{x}_i - \bar{\mathbf{x}})^T. \quad (3)$$

Note that the standard technique (e.g., in Matlab) for sparse data is to compute (3) only for those data pairs where both values $(\mathbf{x}_i)_j$ and $(\mathbf{x}_i)_k$ exist. By setting $\mathbf{v}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, we have for the quadratic form, with an arbitrary vector $\mathbf{x} \neq 0$:

$$\mathbf{x}^T \Sigma \mathbf{x} = \mathbf{x}^T \left[\mathbf{v}_1\, \mathbf{v}_1^T + \ldots + \mathbf{v}_N\, \mathbf{v}_N^T\right] \mathbf{x} = (\mathbf{x}^T \mathbf{v}_1)^2 + \ldots + (\mathbf{x}^T \mathbf{v}_N)^2 \geq 0. \quad (4)$$

This shows that any matrix of the form of (3) is always at least positive semidefinite, with positive eigenvalues if $\mathbf{v}_i$'s span $\mathbf{R}^n$, that is, if $\mathrm{rank}[\mathbf{v}_1 \;\ldots \mathbf{v}_N] \geq n$. The existence of missing values clearly increases the possibility of semidefiniteness.

Now, let $\{\lambda_k, \mathbf{u}_k\}$ be the kth eigenvalue and eigenvector of Σ satisfying

$$\Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad k = 1, \ldots, n. \quad (5)$$

This identity can be written in the matrix form as $\Sigma \mathbf{U} = \mathbf{U}\mathbf{D}$, where $\mathbf{D} = \mathrm{Diag}\{\lambda_1, \ldots, \lambda_n\}$ (vector $\boldsymbol{\lambda}$ as the diagonal matrix) and $\mathbf{U} = \left[\mathbf{u}_1\, \mathbf{u}_2 \ldots \mathbf{u}_n\right]$. Using (5) shows that (2) reduces to $\mathcal{J} = \frac{1}{2} \sum_{k=m+1}^{n} \lambda_k$. This means that the reduced representation consists of those m eigenvectors that correspond to the m largest eigenvalues of matrix Σ. For the unbiased estimate of the sample covariance matrix $\Sigma \simeq \frac{1}{N-1} \Sigma$, one can use scaling such as in (3) because it does not affect eigenvectors or the relative sizes of the eigenvalues. Finally, for any $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} = \mathbf{U}^T \mathbf{x}$, we have

$$\mathbf{x}^T \Sigma \mathbf{x} = \mathbf{y}^T \mathbf{D} \mathbf{y} = \sum_{k=1}^{n} \lambda_k \mathbf{y}_k^2 = \sum_{k=1}^{n} \frac{\mathbf{y}_k^2}{\left(\lambda_k^{-\frac{1}{2}}\right)^2}. \quad (6)$$

Geometrically, this means that in the transformed coordinate system $\mathbf{U}^T\mathbf{e}_k$ ($\mathbf{e}_k$s are the base vectors for the original coordinates), the data define an n-dimensional hyperellipsoid for which the lengths of the principal semi-axis are proportional to $\sqrt{\lambda_k}$.

To this end, we redefine the well-known principle (see, e.g., [15]) for choosing a certain number of principal components in dimension reduction. Namely, the derivations above show that eigenvalues of the sample covariance matrix Σ represent *the variance* along the new coordinate system, $\lambda_k = \sigma_k^2$, whereas the geometric interpretation related to (6) proposes to use the standard deviation $\sigma_k = \sqrt{\lambda_k}$ to assess the variability of data.

Proposition 1. *The relative importance RI_k (in percentages) of a new variable y_k for the principal component transformation based on the sample covariance matrix is defined as $RI_k = 100\frac{\sqrt{\lambda_k}}{\sum_{i=1}^n \sqrt{\lambda_i}}$, where λ_k satisfy* (5). *We refer to $\sqrt{\lambda_i}$ as the estimated variability of the ith (new) variable.*

2.2 Derivation of robust PCA for sparse data

Formally, a straightforward derivation of the classical PCA as given above is obtained from the optimality condition for the least-squares problem (1). Namely, assume that instead of the reduced representation, the problem $\min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$ as in (1) is used to estimate the location of the given data $\{\mathbf{x}_i\}$. In second-order statistics, this provides the sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$, whose explicit formula can be obtained from the optimality condition (see [16]):

$$\frac{d\mathcal{J}(\bar{\mathbf{x}})}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}\frac{1}{2}\sum_{i=1}^N \|\mathbf{x}_i - \mathbf{x}\|^2 = \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{0}. \tag{7}$$

The covariate form of this optimality condition $\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ readily provides us the sample covariance matrix up to the constant $\frac{1}{N-1}$.

Next we assume that there are missing values in the given data. To define their pattern, let us introduce the projection vectors $\mathbf{p}_i$, with $i = 1\ldots, N$ (see [17, 2, 24, 25]), which capture the availability of the components:

$$(\mathbf{p}_i)_j = \begin{cases} 1, \text{if } (\mathbf{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases} \tag{8}$$

We also define the corresponding matrix $\mathbf{P} \in \mathbf{R}^{N\times n}$ that contains these projections in the rows, being of compatible size with the data matrix $\mathbf{X}$.

The spatial median $\mathbf{s}$ with the so-called available data strategy can be obtained as the solution of the projected Weber problem

$$\min_{\mathbf{v}\in\mathbf{R}^n} \mathcal{J}(\mathbf{v}), \quad \text{where} \;\; \mathcal{J}(\mathbf{v}) = \sum_{i=1}^{n_j} \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{v})\|. \tag{9}$$

As described in [16], this optimization problem is nonsmooth, that is, it is not classically differentiable at zero. Instead, the so-called subgradient of $\mathcal{J}(\mathbf{v})$ always exists and is characterized by the condition

$$\partial \mathcal{J}(\mathbf{v}) = \sum_{i=1}^{N} \boldsymbol{\xi}_i \text{ for } \begin{cases} (\boldsymbol{\xi}_i)_j = & \dfrac{\text{Diag}\{\mathbf{p}_i\}(\mathbf{v}-\mathbf{x}_i)_j}{\|\text{Diag}\{\mathbf{p}_i\}(\mathbf{v}-\mathbf{x}_i)\|}, \text{if } \|\text{Diag}\{\mathbf{p}_i\}(\mathbf{u}-\mathbf{x}_i)\| \neq 0, \\ \|\boldsymbol{\xi}_i\| \leq & 1, \text{ when } \|\text{Diag}\{\mathbf{p}_i\}(\mathbf{u}-\mathbf{x}_i)\| = 0. \end{cases} \tag{10}$$

Then, the minimizer $\mathbf{s}$ of (9) satisfies $\mathbf{0} \in \partial \mathcal{J}(\mathbf{s})$. In [20] it is shown, for the complete data case, that if the sample $\{\mathbf{x}_i\}$ belongs to a Euclidean space and is not concentrated on a line, the spatial median $\mathbf{s}$ is unique. In practice (see [2]), one can obtain an accurate approximation for the solution of the nonsmooth problem by solving the following equation corresponding to the regularized form

$$\sum_{i=1}^{N} \frac{\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\max\{\|\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|, \varepsilon\}} = \mathbf{0} \quad \text{for } \varepsilon > 0. \tag{11}$$

This can be solved using the SOR (Sequential Overrelaxation) algorithm [2] with the overrelaxation parameter $\omega = 1.5$. For simplicity, define $\|\mathbf{v}\|_\varepsilon = \max\{\|\mathbf{v}\|, \varepsilon\}$.

To this end, the comparison of (7) and (11) allows us to define the *robust covariance matrix* corresponding to the spatial median s:

$$\Sigma_R = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon} \right) \left(\frac{\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\text{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon} \right)^T. \tag{12}$$

This form can be referred to as the *multivariate sign covariance matrix* [5, 28, 7]. By construction, the nonzero covariate vectors have a unit length, so that they only accumulate the deviations of angles and not the sizes of the available variables. Such an observation is related to one perspective on statistical robustness that can be formalized using the so-called influence function [9]. Using Σ_R as the sample covariance matrix, one can, by again solving the corresponding eigenvalue problem (5), recover a new basis $\{\mathbf{u}_k\}$ for which the corresponding eigenvalues $\{\lambda_k\}$, again, explain the amount of variability along the new coordinates. Because Σ_R is based on the first-order approximation, the nonnegative eigenvalues readily correspond to the geometric variability represented by the standard deviation in the second-order statistics, and, then, we do not need to take any square roots when computing the relative importances of the robust procedure as in Proposition 1. Hence, the two PCA approaches are comparable to each other.

2.3 Projection using PCA-based transformation

In the matrix form, the existence of a new basis in the columns of the given unitary matrix $\mathbf{U}$, and given a complete location estimate for the sparse data $\mathbf{s} \in \mathbf{R}^n$ (i.e., the spatial median), for which we define the corresponding matrix $\mathbf{S} \in \mathbf{R}^{N \times n}$ by replication of $\mathbf{s}^T$ in N rows, yields the transformed data matrix

$$\mathbf{Y} = (\mathbf{P} \circ (\mathbf{X} - \mathbf{S}))\,\mathbf{U}, \tag{13}$$

where $\circ$ denotes the Hadamard product. When $\mathbf{U}$ is ordered based on RI_k's, the dimension reduction is obtained by selecting only m of the n coordinates (columns) in $\mathbf{Y}$. Hence, we see that even if there are missing values in the original data, the resulting new data vectors become complete. We also know from the basic linear algebra that, for complete data, both the length of the original vectors and the angle between any two vectors are preserved in (13) because $\mathbf{U}$ is unitary. However, in the case of missing data, some of the coordinate values of the original vectors are not present, and then, presumably, the transformed vectors in $\mathbf{Y}$ are of smaller length, i.e., closer to the origin in the transformed space. Moreover, the angles might also become degraded. These simple observations readily raise some doubts concerning the available data strategy in the form of incomplete data vectors as proposed in (12).

2.4 Two modifications of the robust PCA procedure

Let us define two modifications of the robust PCA procedure that are based on the similar form of the covariance matrix as defined in (12). As discussed above, both the amount of variability of data and/or the main directions of variability might be underestimated due to sparse data vectors, that is, missing coordinate values. Our suggested modifications are both based on a simple idea: use only the "almost complete" data in estimation (cf. the cascadic initializations of robust clustering in [24, 25]). Note that this is one step further than the typical way of using only the complete pairs or complete observations in the computation of a covariance matrix.

The first suggested modification, for the computation of the relative importances of the principal components, is related to using the actual projections along the new coordinate axis for this purpose. Similar to the alpha-trimmed mean [3], which presumably neglects outlying observations, we use (see the tests in [26]) the 10% and 90% percentiles, denoted as $\mathrm{prc}_{10}(\cdot)$ and $\mathrm{prc}_{90}(\cdot)$, related to the transformed data matrix $\mathbf{Y}$ in (13). Namely, for the each new variable $\{y_k\}$, its estimated variability is computed as

$$RI_k = 100(\mathrm{prc}_{90}(\{y_k\}) - \mathrm{prc}_{10}(\{y_k\})). \tag{14}$$

Moreover, because it is precisely the sparsity that diminishes the lengths and angles of the transformed data vectors, we restrict the computation of (14) to that subset of the original data, where at most one variable is missing from an observation $\mathbf{x}_i$. This subset satisfies $\sum_{j=1}^{n}(\mathbf{p}_i)_j \geq n-1$.

Our second suggested modification uses a similar approach, but already directly for the robust covariance matrix (12), by taking into account only those observations of which at most one variable is missing. Hence, we define the following subsets of the original set of indices $\mathcal{N} = \{1, 2, \ldots, N\}$:

$$\begin{aligned} I_c &= \{i \in \mathcal{N} \mid \mathbf{x}_i \text{ is complete}\}, \\ I_j &= \{i \in \mathcal{N} \mid \text{variable j is missing from } \mathbf{x}_\mathrm{i}\}. \end{aligned}$$

We propose computing a reduced, robust covariance matrix $\widetilde{\Sigma}_R$ as

$$\widetilde{\Sigma}_R = \frac{1}{\widetilde{N}-1}\left(\sum_{i\in I_c}\mathbf{v}_i\mathbf{v}_i^T + \sum_{j=1}^{n}\sum_{i\in I_j}\mathbf{v}_i\mathbf{v}_i^T\right), \quad \mathbf{v}_i = \frac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon},$$

with $\widetilde{N} = \ |I_c\ | + \sum_j |I_j|$. Hence, only that part of the first-order covariability that corresponds to the almost complete observations is used.

3 Computational results

Computational experiments in the form of simulated test cases, when knowing the target result, are given first. The parametrized test is introduced in Section 3.1, and the computational results for the different procedures are provided in Section 3.2. Finally, we apply the best methods to analyze the educational data of PISA in Section 3.3. As a reference method related to the classical, second-order statistics as derived in Section 2.1, with sparse data, we use the Matlab's PCA routine with the '*pairwise*' option.

3.1 The simulated test cases

For simplicity, we fix the number of observations as $N = 1000$. For the fixed size of an observation n, let us define a vector of predetermined standard deviations as $\boldsymbol{\sigma} = \begin{bmatrix}\sigma_1 \ \sigma_2 \ \dots \ \sigma_n\end{bmatrix}$. Moreover, let $\mathbf{R}_{a,b}(\theta) \in \mathbf{R}^{n\times n}$ be an orthonormal (clockwise) rotation matrix of the form

$$\mathbf{R}_{ab}(\theta) = \{\mathbf{M} = \mathbf{I}_n \wedge \mathbf{M}_{aa} = \mathbf{M}_{bb} = \cos(\theta),\ \mathbf{M}_{ab} = -\mathbf{M}_{ba} = -\sin(\theta)\},$$

where $\mathbf{I}_n$ denotes the $n\times n$ identity matrix. Then, the simulated data $\{\mathbf{d}_i\}_{i=1}^N$ is generated as

$$\begin{aligned}\mathbf{d}_i^T \sim & \frac{\boldsymbol{\sigma}}{2} + \left[\mathcal{N}(0,\sigma_1)\ \mathcal{N}(0,\sigma_2)\ \dots\ \mathcal{N}(0,\sigma_n)\right] \\ & + \eta_i\left[\mathbf{R}_n\left[\mathcal{U}([-\sigma_1,\sigma_1])\ \mathcal{U}([-\sigma_2,\sigma_2])\ \dots\ \mathcal{U}([-\sigma_n,\sigma_n])\right]^T\right]^T,\end{aligned} \tag{15}$$

where $\mathcal{N}(0,\sigma)$ denotes the zero-mean normal distribution with standard deviation σ and $\mathcal{U}([-c,c])$ the uniform distribution on the interval $[-c,c]$, respectively. $\mathbf{R}_n$ defines the n-dimensional rotation that we use to orientate the latter noise term in (15) along the diagonal of the hypercube, that is, we always choose $\theta = \frac{\pi}{4}$ and take, for the actual tests in $2D, 3D, 4D$, and $6D$,

$$\begin{aligned}&\mathbf{R}_2 = \mathbf{R}_{12}(\theta),\quad \mathbf{R}_3 = \mathbf{R}_{23}(\theta)\mathbf{R}_{12}(\theta),\quad \mathbf{R}_4 = \mathbf{R}_{14}(\theta)\mathbf{R}_{23}(\theta)\mathbf{R}_{34}(\theta)\mathbf{R}_{12}(\theta),\\ &\mathbf{R}_6 = \mathbf{R}_{36}(\theta)\mathbf{R}_{45}(\theta)\mathbf{R}_{56}(\theta)\mathbf{R}_{14}(\theta)\mathbf{R}_{23}(\theta)\mathbf{R}_{34}(\theta)\mathbf{R}_{12}(\theta).\end{aligned}$$

Finally, a random sparsity pattern of a given percentage of missing values represented by the matrix $\mathbf{P}$ as defined in (8) is attached to data.

To conclude, the simulated data are parametrized by the vector $\boldsymbol{\sigma}$, which defines the true data variability. Moreover, the target directions of the principal components are just the original unit vectors $\mathbf{e}_k,\ k = 1, \ldots, n$. Their estimation is disturbed by the noise, which comes from the uniform distribution whose width coordinatewise coincides with the clean data. Because the noise is rotated towards the diagonal of the hypercube, its maximal effect is characterized by $\frac{\max_k \sigma_k}{\min_k \sigma_k}$. By choosing σ_k's as the powers of two and three for $n = 2, 3, 4, 6$, we are then gradually increasing the effect of the error when the dimension of the data is increasing. Finally, we fix the amount of noise to 10% so that $\eta_i = 1$ with a probability of 0.1 in (15). In this way, testing up to 40% of missing values randomly attached to $\{\mathbf{d}_i\}$ will always contain less than 50% of the degradations (missing values and/or noise) as a whole.

3.2 Results for the simulated tests

The test data generation was repeated 10 times, and the means and standard deviations (in parentheses) over these are reported. As the error measure for the directions of $\{\mathbf{u}_k\}$, we use their deviation from being parallel to the target unit vectors. Hence, we take $\text{DirE} = \max_k\{1 - |\mathbf{u}_k^T \mathbf{e}_k|\},\ k = 1, \ldots, n$, such that $\text{DirE} \in [0, 1]$. In the result tables below, we report the relative importances of RI_k in the order of their importance. 'Clas' refers to the classical PCA, 'Rob' to the original robust formulation, 'RobP' to the modification using percentiles for the importances, and 'RobR' to the use of the reduced covariance matrix $\tilde{\Sigma}_R$. The real relative importances ('True') by generation are provided in the third column.

From all simulated tests (Tables 1-4), we see that the the classical method and 'RobP' show the closest relative importances of the principal components to the true geometric variability in the data. Moreover, both of these approaches show a very stable behavior, and the results for the relative importances do not change that much, even when a high number of missing data is present. The results for the other two approaches, the basic robust and 'RobR', on the other hand, are much less stable, and particularly the basic robust procedure starts to underestimate the relative importances of the major components when the amount of missing data increases.

The directions remain stable for all the simulated test cases, even when a large amount of missing data is present. Over all the simulated tests, the 'RobP' with the original robust covariance bears the closest resemblance to the true directions. It can tolerate more noise compared to 'Clas', as shown in Table 3. We also conclude that the missing data do not affect the results of the PCA procedures as much as the noise. Tables 3 and 4 show that, for a large noise, the increase in sparsity can actually improve the performance of the robust method because it decreases the absolute number of noisy observations. Interestingly, as can be seen from Table 4, the geometric variability was estimated accurately, even if the directions were wrong.

Table 1. Results for $\boldsymbol{\sigma} = [3\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	75.0(0.00)	73.7(0.8)	73.0(1.1)	73.0(1.3)	73.0(1.1)
	2	25.0(0.00)	26.3(0.8)	27.0(1.1)	27.0(1.3)	27.0(1.1)
	DirE	-	0.001	0.004		0.004
10%	1	75.0(0.00)	73.9(0.9)	68.9(1.2)	73.2(1.3)	68.9(1.2)
	2	25.0(0.00)	26.1(0.9)	31.1(1.2)	26.8(1.3)	31.1(1.2)
	DirE	-	0.001	0.005		0.005
20%	1	75.0(0.00)	73.5(1.2)	65.1(1.0)	72.5(1.7)	65.1(1.0)
	2	25.0(0.00)	26.5(1.2)	34.9(1.0)	27.5(1.7)	34.9(1.0)
	DirE	-	0.001	0.009		0.009
30%	1	75.0(0.00)	73.8(1.0)	62.4(0.9)	73.0(1.4)	62.4(0.9)
	2	25.0(0.00)	26.2(1.0)	37.6(0.9)	27.0(1.4)	37.6(0.9)
	DirE	-	0.001	0.003		0.003
40%	1	75.0(0.00)	74.0(0.8)	60.3(1.6)	73.1(1.4)	60.3(1.6)
	2	25.0(0.00)	26.0(0.8)	39.7(1.6)	26.9(1.4)	39.7(1.6)
	DirE	-	0.002	0.008		0.008

3.3 Results for PISA data set

Next, we apply the different PCA methods tested in the previous section to a large educational data set, namely the latest data from the Programme for International Student Assessment[1] (PISA 2012). The data contain 485490 observations, and as variables we use the 15 scale indices [24] that are known to explain the student performance in mathematics, the main assessment area in PISA 2012. The scale indices are derived variables that summarize information from student background questionnaires [22], and are scaled so that their mean is zero with a standard deviation of one. Due to the rotated design of PISA (each student answers only one of the three different background questionnaires), this data set has 33.24% of missing data by design, a special case of MCAR.

In Table 5, the relative importances $\{RI_k\}$ are depicted. The table also shows the variance-based view for the classical method, denoted as 'ClsVar'. As can be seen from the table, the first principal component is much higher for 'ClsVar' than for the other approaches. In consequence, fewer principal components would be selected with 'ClsVar' when a certain threshold of how much the principal components should account for is given. As illustrated in Fig. 1, if the threshold is set to 90%, we would select 11 components with 'ClsVar' but 13 for both the classical PCA and for the 'RobP'.

In Fig. 2, the loadings of the first two principal components are visualized for the classical and for the robust version. We see that for both versions, the three scale indices ANXMAT, FAILMAT, and ESCS are the most distinct from the others. However, the robust version is able to distinguish this finding more clearly. That *index of economic, social and cultural status* (ESCS) accounts for

[1] Available at `http://www.oecd.org/pisa/pisaproducts/`.

Table 2. Results for $\boldsymbol{\sigma} = [4\ 2\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	57.1(0.00)	56.3(0.7)	58.6(1.3)	55.8(1.0)	58.6(1.3)
	2	28.6(0.00)	28.6(0.8)	28.9(1.2)	28.7(1.0)	28.9(1.2)
	3	14.3(0.00)	15.2(0.3)	12.6(0.4)	15.5(0.4)	12.6(0.4)
	DirE	-	0.005	0.017		0.017
10%	1	57.1(0.00)	56.3(0.8)	55.7(1.3)	55.8(1.0)	54.5(1.6)
	2	28.6(0.00)	28.6(0.9)	30.0(1.2)	28.6(0.9)	30.7(1.2)
	3	14.3(0.00)	15.1(0.4)	14.3(0.6)	15.5(0.5)	14.8(0.8)
	DirE	-	0.008	0.015		0.015
20%	1	57.1(0.00)	56.2(0.8)	51.7(1.4)	55.6(1.1)	51.7(1.4)
	2	28.6(0.00)	28.6(0.9)	30.7(1.2)	28.8(1.3)	31.5(1.5)
	3	14.3(0.00)	15.2(0.3)	17.6(0.8)	15.6(0.5)	16.7(0.7)
	DirE	-	0.005	0.020		0.014
30%	1	57.1(0.00)	56.0(0.7)	49.2(0.7)	55.3(0.9)	50.9(0.7)
	2	28.6(0.00)	28.8(0.8)	31.6(0.8)	29.0(0.9)	32.1(1.3)
	3	14.3(0.00)	15.2(0.4)	19.2(1.0)	15.7(0.5)	17.1(1.3)
	DirE	-	0.006	0.013		0.012
40%	1	57.1(0.00)	56.2(0.9)	46.2(1.4)	55.8(1.2)	49.9(1.6)
	2	28.6(0.00)	28.7(1.2)	32.0(1.7)	28.5(1.3)	32.5(1.7)
	3	14.3(0.00)	15.1(0.4)	21.8(1.1)	15.6(0.7)	17.6(1.0)
	DirE	-	0.010	0.014		0.013

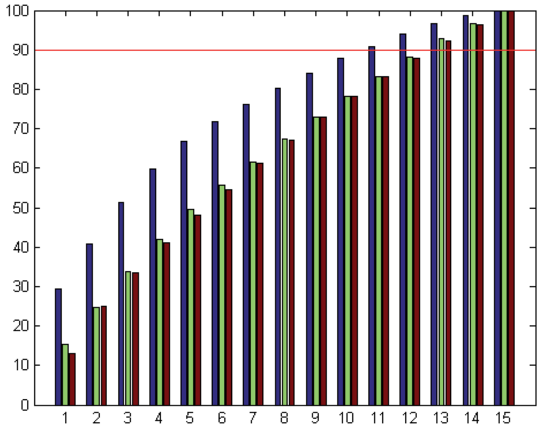


Fig. 1. Cumulative sum of the relative importances for the classical PCA using variance, the classical PCA, and the robust PCA using percentiles (from left to right).

much of the variability in the data, being the "strongest single factor associated with performance in PISA" [21], is always highlighted in PISA documentations and can be clearly seen in Fig. 2, especially from the robust PC 1.

Table 3. Results for $\boldsymbol{\sigma} = [27\ 9\ 3\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	67.5(0.00)	62.5(0.8)	66.9(0.9)	63.6(1.1)	66.9(0.9)
	2	22.5(0.00)	22.1(0.6)	23.6(0.8)	22.9(0.8)	23.6(0.8)
	3	7.5(0.00)	10.1(0.2)	6.9(0.4)	9.1(0.4)	6.9(0.4)
	4	2.5(0.00)	5.3(0.2)	2.6(0.2)	4.5(0.2)	2.6(0.2)
	DirE	-	0.168	0.080		0.080
10%	1	67.5(0.00)	62.5(0.8)	62.3(1.3)	64.0(1.2)	60.3(2.3)
	2	22.5(0.00)	22.1(0.6)	25.7(0.9)	23.0(0.9)	26.9(1.7)
	3	7.5(0.00)	10.0(0.2)	8.6(0.5)	9.0(0.4)	9.2(0.6)
	4	2.5(0.00)	5.3(0.2)	3.5(0.3)	4.0(0.2)	3.6(0.4)
	DirE	-	0.157	0.045		0.046
20%	1	67.5(0.00)	62.5(1.0)	56.9(1.2)	64.2(1.2)	58.3(1.7)
	2	22.5(0.00)	22.1(0.7)	27.4(1.0)	22.9(0.9)	28.1(1.1)
	3	7.5(0.00)	10.1(0.3)	11.0(0.6)	8.9(0.5)	9.8(1.0)
	4	2.5(0.00)	5.4(0.3)	4.7(0.4)	3.9(0.2)	3.7(0.4)
	DirE	-	0.164	0.032		0.031
30%	1	67.5(0.00)	62.7(0.8)	52.1(1.6)	64.4(0.9)	57.7(1.3)
	2	22.5(0.00)	22.0(0.6)	28.2(1.4)	23.2(0.6)	28.2(1.5)
	3	7.5(0.00)	10.0(0.4)	13.2(1.0)	8.6(0.4)	10.2(0.5)
	4	2.5(0.00)	5.3(0.3)	6.5(0.5)	3.8(0.2)	3.9(0.4)
	DirE	-	0.177	0.023		0.038
40%	1	67.5(0.00)	62.7(0.8)	46.9(0.8)	64.2(1.4)	55.9(1.2)
	2	22.5(0.00)	22.2(0.6)	28.7(1.0)	23.5(1.3)	29.8(1.5)
	3	7.5(0.00)	9.9(0.3)	15.7(0.5)	8.8(0.3)	10.8(1.1)
	4	2.5(0.00)	5.2(0.3)	8.6(0.8)	3.6(0.4)	3.5(0.5)
	DirE	-	0.189	0.016		0.040

4 Conclusions

Although PCA is one of the most widely used ML and DM techniques, systematic testing and assessment of PCA in the presence of missing data seem to still be an important topic to study. In this article, we have proposed a robust PCA method and two modifications (one using percentiles for the importance and one with a reduced covariance matrix) of this method. The testing of these three approaches was done in comparison with the classical, reference PCA for sparse data. First, we illustrated the results for carefully designed simulated data and then for a large, real educational data set.

From the simulated tests, we concluded that the percentiles-based robust method and the classical PCA showed the best results, especially when the relative importance of the principal components were compared with the true variability of the data. The basic robust approach started to underestimate the relative importance of the major components when the amount of missing data increased. The results of the simulated tests were stable, and the variance be-

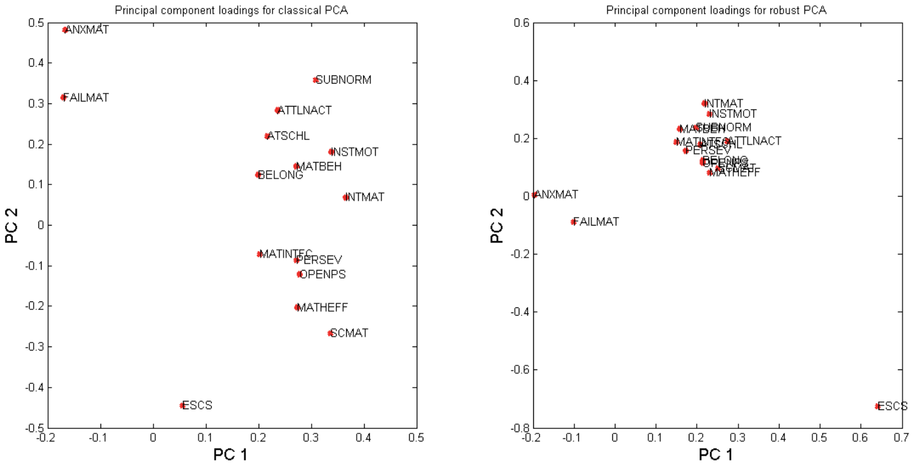


Fig. 2. Principal component loadings for PISA data for the classical (left) and robust (right) approaches.

tween repeated test runs was very small. Likewise, the estimated directions remained also stable even with a large amount of missing data. Tests with PISA data showed that the proposed robust methods are applicable for large, real data sets with one-third of the values missing, where the interpretation of the robust result yielded clearer known discrimination of the original variables compared to the classical PCA.

The classical PCA uses variance to estimate the importance of the principal components, which highlights (as demonstrated in Table 5 and Figure 1) the major components. As shown by the simulated results, it is more prone to nongaussian errors in the data. These points might explain some of the difficulties the classical method faced in applications [23]. In [14], seven distinctions of the PCA problem in the presence of missing values were listed: 1) no analytical solution since even the estimation of the data covariance matrix is nontrivial, 2) the optimized cost function typically has multiple local minima, 3) no analytical solution even for the location estimate, 4) standard approaches can lead to overfitting, 5) algorithms may require heavy computations, 6) the concept of the PCA basis in the principal subspace is not easily generalized, and 7) the choice of the dimensionality of the principal subspace is more difficult than in classical PCA. We conclude that the proposed robust methods successfully addressed all these distinctions: 1) well-defined covariance matrix, 2) being positive semidefinite, 3) a unique location estimate in the form of the spatial median, 4) resistance to noise due to robustness, 5) the same linear algebra as in the classical approach, and 6)–7) a geometrically consistent definition of the principal subspace and its dimension related to the data variability.

Acknowledgments. The authors would like to thank Professor Tuomo Rossi for many helpful discussions on the contents of the paper.

Table 4. Results for $\boldsymbol{\sigma} = [32\ 16\ 8\ 4\ 2\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	50.8(0.00)	48.1(0.5)	55.3(1.0)	48.6(0.8)	55.3(1.0)
	2	25.4(0.00)	24.3(0.4)	26.7(0.6)	24.4(0.4)	26.7(0.6)
	3	12.7(0.00)	12.6(0.2)	10.6(0.4)	12.7(0.3)	10.6(0.4)
	4	6.3(0.00)	7.5(0.2)	4.7(0.3)	7.1(0.2)	4.7(0.3)
	5	3.2(0.00)	4.4(0.1)	1.6(0.1)	4.3(0.1)	1.6(0.1)
	6	1.6(0.00)	3.2(0.1)	1.0(0.1)	2.8(0.2)	1.0(0.1)
	DirE	-	0.298	0.374		0.374
10%	1	50.8(0.00)	48.0(0.6)	51.6(1.1)	48.5(1.1)	51.0(1.9)
	2	25.4(0.00)	24.3(0.5)	27.5(0.8)	24.8(0.6)	28.1(1.1)
	3	12.7(0.00)	12.6(0.2)	12.0(0.5)	12.8(0.4)	12.0(0.9)
	4	6.3(0.00)	7.5(0.2)	5.5(0.2)	7.0(0.2)	5.5(0.4)
	5	3.2(0.00)	4.4(0.1)	2.2(0.2)	4.2(0.1)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	1.3(0.1)	2.7(0.2)	1.3(0.2)
	DirE	-	0.318	0.277		0.358
20%	1	50.8(0.00)	48.2(0.5)	48.6(1.0)	48.9(1.6)	51.3(1.4)
	2	25.4(0.00)	24.2(0.5)	27.3(0.8)	24.6(0.6)	27.3(0.7)
	3	12.7(0.00)	12.7(0.3)	13.2(0.8)	13.0(0.6)	12.3(1.2)
	4	6.3(0.00)	7.4(0.2)	6.4(0.3)	7.0(0.3)	5.5(0.6)
	5	3.2(0.00)	4.4(0.2)	2.7(0.3)	4.0(0.2)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	1.7(0.2)	2.4(0.2)	1.3(0.3)
	DirE	-	0.372	0.090		0.137
30%	1	50.8(0.00)	48.1(0.6)	43.8(1.2)	48.6(1.4)	49.4(2.4)
	2	25.4(0.00)	24.3(0.5)	27.5(0.8)	25.0(0.8)	28.5(1.7)
	3	12.7(0.00)	12.6(0.1)	15.0(0.6)	12.9(0.5)	12.5(0.8)
	4	6.3(0.00)	7.5(0.2)	7.6(0.5)	7.1(0.4)	5.8(0.8)
	5	3.2(0.00)	4.3(0.1)	3.8(0.5)	4.0(0.2)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	2.3(0.3)	2.4(0.2)	1.5(0.4)
	DirE	-	0.335	0.092		0.468
40%	1	50.8(0.00)	48.0(0.6)	39.7(1.5)	48.3(1.7)	50.2(2.9)
	2	25.4(0.00)	24.3(0.4)	26.6(1.0)	25.1(1.1)	28.3(2.4)
	3	12.7(0.00)	12.6(0.3)	15.8(1.0)	13.0(0.7)	11.9(1.3)
	4	6.3(0.00)	7.5(0.2)	9.5(0.8)	7.3(0.3)	6.0(0.8)
	5	3.2(0.00)	4.4(0.3)	5.1(0.5)	3.9(0.3)	2.2(0.3)
	6	1.6(0.00)	3.1(0.2)	3.3(0.4)	2.3(0.2)	1.3(0.2)
	DirE	-	0.516	0.078		0.518

Table 5. Results for PISA data

	RI_1	RI_2	RI_3	RI_4	RI_5	RI_6	RI_7	RI_8	RI_9	RI_{10}	RI_{11}	RI_{12}	RI_{13}	RI_{14}	RI_{15}
ClsVar	29.5	11.4	10.4	8.6	6.8	5.0	4.4	4.1	3.8	3.7	3.2	3.0	2.8	2.0	1.3
Cls	15.3	9.5	9.1	8.3	7.3	6.3	5.9	5.7	5.5	5.4	5.0	4.8	4.7	4.0	3.3
RobP	13.1	11.9	8.6	7.5	7.2	6.5	6.5	5.9	5.9	5.2	4.8	4.8	4.5	3.9	3.7

Bibliography

[1] E. Alpaydin. *Introduction to Machine Learning.* The MIT Press, Cambridge, MA, USA, 2nd edition, 2010.

[2] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing.* University of Jyväskylä, 2006.

[3] J. Bednar and T. Watt. Alpha-trimmed means and their relationship to median filters. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(1):145–153, 1984.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[5] C. Croux, E. Ollila, and H. Oja. Sign and rank covariance matrices: statistical properties and application to principal components analysis. In *Statistical data analysis based on the L1-norm and related methods*, pages 257–269. Springer, 2002.

[6] A. d'Aspremont, F. Bach, and L. E. Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

[7] D. Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.).* Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[9] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[10] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[11] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods.* Edward Arnold, London, 1998.

[12] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[13] P. J. Huber. *Robust Statistics.* John Wiley & Sons Inc., New York, 1981.

[14] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

[15] I. Jolliffe. *Principal component analysis.* Wiley Online Library, 2005.

[16] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[17] T. Kärkkäinen and J. Toivanen. Building blocks for odd-even multigrid with applications to reduced systems. *Journal of Computational and Applied Mathematics*, 131:15–33, 2001.

[18] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*, volume 4. Wiley New York, 1987.

[19] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[20] P. Milasevic and G. R. Ducharme. Uniqueness of the spatial median. *Ann. Statist.*, 15(3):1332–1333, 1987.

[21] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[22] OECD. *PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III)*. PISA, OECD Publishing, 2013.

[23] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. In *ACM SIGMETRICS Performance Evaluation Review*, volume 35, pages 109–120. ACM, 2007.

[24] M. Saarela and T. Kärkkäinen. Discovering gender-specific knowledge from Finnish basic education using PISA scale indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[25] M. Saarela and T. Kärkkäinen. Analysing student performance using sparse data of core bachelor courses. *To appear in JEDM-Journal of Educational Data Mining*, 2015.

[26] S. M. Stigler. Do robust estimators work with real data? *The Annals of Statistics*, pages 1055–1098, 1977.

[27] J. R. Van Ginkel, P. M. Kroonenberg, and H. A. Kiers. Missing data in principal component analysis of questionnaire data: a comparison of methods. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–18, 2013.

[28] S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575, 2000.

PVIII

PREDICTING MATH PERFORMANCE FROM RAW LARGE-SCALE EDUCATIONAL ASSESSMENTS DATA: A MACHINE LEARNING APPROACH

by

Mirka Saarela, Bülent Yener, Mohammed Zaki, Tommi Kärkkäinen 2016

Machine Learning for Digital Education and Assessment Systems Workshop of
the 33rd International Conference on Machine Learning, pp. 1–8

Predicting Math Performance from Raw Large-Scale Educational Assessments Data: A Machine Learning Approach

Mirka Saarela[*,†] MIRKA.SAARELA@JYU.FI
Bülent Yener[†] YENER@CS.RPI.EDU
Mohammed J. Zaki[†] ZAKI@CS.RPI.EDU
Tommi Kärkkäinen[*] TOMMI.KARKKAINEN@JYU.FI
[*]University of Jyväskyla, Department of Mathematical Information Technology, 40014 Jyväskylä, Finland
[†]Rensselaer Polytechnic Institute, Computer Science Department, 12180 Troy, New York, USA

Abstract

Large-scale educational assessment studies (LSAs) regularly collect massive amounts of very rich cognitive and contextual data of whole student populations. Currently, LSAs are limited to reporting student proficiencies in the form of *plausible values* (PVs). PVs are random draws from the posterior distribution of a student's ability, which is based on the Bayesian approach with the prior distribution modeling the student background within the population and the likelihood test item response using the Rasch model. While PVs have shown to be a reliable estimate for proficiencies of populations, a more comprehensive study of these rich data sets by deploying machine learning algorithms may provide a better understanding of the underlying factors affecting student performance and thus yield to better and more interpretable predictive models. This paper presents such a novel approach to learn directly from LSA data by deploying a combination of both unsupervised and supervised learning feature selection algorithms to predict student performance on math scores. Our technique learns the difficulty level of different math questions and predicts weather or not a student with a particular background profile will be successful in answering correctly.

1. Introduction

Since 2000 triennially, the Organisation for Economic Co-operation and Development (OECD) collects a massive amount of data of stratified samples of 15-year-old students from all over the world for the Programme for International Student Assessment (PISA). The sampled students not only take a cognitive test—in which they have to demonstrate their math, reading and science skills—but also reply to a questionnaire, in which they provide information about their social and economical background, as well as their motivations, behaviors, and attitudes towards various aspects of education. All collected data is publicly available[1] and according to the OECD, of very high quality in terms of degree of validity and reliability (OECD, 2009; 2012). Moreover, these data are comparable throughout different countries so that they provide a very rich database for educational machine learning (ML) and data mining (DM) applications.

The participating countries pay large sums of money (Musik, 2016) primarily with the goal to utilize PISA data and analysis results for research. However, as concluded by Rutkowski et al. (2010), not many researchers work with these freely available and high quality datasets because of the many technical complexities within them. The major difficulty of conducting secondary analysis with PISA data is that many desired properties that describe the students are not originally observed features, but are already preprocessed and made available as derived variables through a combination of different state-of-the-art methodologies. One example is that there are no single performance scores for the cognitive test in PISA datasets. Instead, for each student and each assessment domain—reading, math, and science—five plausible values (PVs) are reported.

The PVs are random draws from the posterior distribution of a student's ability, which is defined as

$$f(\beta \mid x_i, y_i) \propto P(x_i \mid \beta, \delta) f(\beta \mid \lambda, y_i), \qquad (1)$$

MLDEAS workshop papers of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[1] PISA data can be downloaded from `http://www.oecd.org/pisa/pisaproducts/`.

where $P(x_i \mid \beta, \delta)$ denotes a Rasch Model (Rasch, 1960) given the student's ability β and the test items' difficulties δ, and $f(\beta \mid \lambda, y_i)$ denotes a population model with the background information of the student encoded in y_i[2]. This population model for a student i is estimated with the latent regression model (Tarpey & Petkova, 2010) $\beta_i = y_i^T \lambda + \epsilon_i$, where $\epsilon_i = \mathcal{N}(0, \sigma^2)$ (Marsman, 2014; OECD, 2014), and with λ denoting the regression coefficients.

PVs have shown to be a reliable estimate for proficiencies of populations (Monseur & Adams, 2008; Wu & Adams, 2002; OECD, 2009) and are used not only in PISA, but also in other LSA studies, such as the National Assessment of Educational Progress[3], the European Survey on Language Competences[4], the Trends in International Mathematics and Science Study, and the Progress in International Reading Literacy Study[5]. However, these estimations are done on normalized data and are based on linear regression (i.e., the λ parameter in f above). Thus, it is worth investigating how deploying a general framework of ML can complement the current state of art by using the raw data which is publicly available.

In this paper, we describe a ML approach that combines unsupervised learning with several supervised learning algorithms and deploys various feature selection algorithms by working directly with raw data. One particular challenge is the sparsity of raw cognitive data due to the design of tests, and missing values in the questionnaire data (Saarela & Kärkkäinen, 2014; 2015a;b; Kärkkäinen & Saarela, 2015; Rutkowski et al., 2010). This work addresses the high sparsity of the cognitive data by clustering the scored cognitive item response data into several difficulty bins and using each bin as a label as we explain later in Section 3.1. Since there were enough data points without missing contextual data from the PISA background questionnaire, we defer to imputation for the future work and focused on complete data. We examined the interaction between different classifier-feature selection algorithms and show that ML is a promising and complementary approach to understand and predict student performance.

The structure of this paper is as follows. In Section 2, we describe the PISA data. After that, our overall method is explained in Section 3, and the experimental results are presented in Section 4. Finally, in Section 5, overall results are summarized and directions for further work are discussed.

Table 1. Item cluster allocation to booklets in PISA 2012. PM denotes cluster of math, PR cluster of reading, and PS cluster of science items.

BOOKLET ID	ITEM CLUSTER			
B1	PM5	PS3	PM6A	PS2
B2	PS3	PR3	PM7A	PR2
B3	PR3	PM6A	PS1	PM3
B4	PM6A	PM7A	PR1	PM4
B5	PM7A	PS1	PM1	PM5
B6	PM1	PM2	PR2	PM6A
B7	PM2	PS2	PM3	PM7A
B8	PS2	PR2	PM4	PS1
B9	PR2	PM3	PM5	PR1
B10	PM3	PM4	PS3	PM1
B11	PM4	PM5	PR3	PM2
B12	PS1	PR1	PM2	PS3
B13	PR1	PM1	PS2	PR3

2. Data

We use the two main student datasets from the latest PISA assessment, which was conducted in 2012 (the 2015 data is not yet public): the *scored cognitive item response* and the *student questionnaire data file*. Both datasets have 485,490 observations (the students who attended the 2012 PISA assessment) and a couple of hundreds of variables.

As explained above, every student that attends the PISA test is assigned only a small fraction of the whole item battery. In PISA 2012, there were 13 main different tests—called *booklets*—and 210 different cognitive test items. Since mathematics was the main assessment domain in PISA 2012, the majority of the items, i.e. 108 of them, are items that test the math proficiency of the students. These cognitive test items were organized into groups—in PISA denoted as *item clusters*—so that each booklet contained four item clusters (this is illustrated in Table 1) and was estimated to be completable in two hours. As can be seen from Table 1, each booklet contained at least one cluster with math items. Our goal in this study is to predict the math performance of the students, which is why we use the sparse $108 \times 485,490$ matrix of the scored math items for building the labels of our classifiers (this will be further explained in Section 3.1).

For the classification features, we are interested in all attributes that are directly concerned with the students' attitudes towards mathematics and that might explain their math performance. In the PISA background questionnaire, there are 53 different math attitudinal statement questions[6], in each of which the student is asked to tick one box of a Likert-scale depending on the degree to which he or she agrees (*totally disagree*, *disagree*, *agree*, or *totally agree*)

[2] In the official PISA literature, it is not explicitly reported which features of the student's background are actually taken into account (OECD, 2014). However, Monseur and Adams (2008) argue that all information from the background questionnaire is utilized.

[3] `nces.ed.gov/nationsreportcard/`

[4] `www.surveylang.org/`

[5] See both `http://timssandpirls.bc.edu/`

[6] Variables ST29Q01–ST46Q09 (position 67–119) in PISA questionnaire data set, see `https://www.oecd.org/pisa/pisaproducts/PISA12_stu_codebook.pdf`

with the given statement. Examples of such statements include *I will learn many things in mathematics that will help me get a job* and *my parents believe studying mathematics is important*. All 53 questions/statements can be found in Figure 1. We select all students that have non-missing values for all of these questions. Because of the rotated design in PISA, these are a bit less than one third of the students from each country. For example, in the Finnish subset, there are 2,491 (out of 8,829) students, which have non-missing values for all these 53 features, and in the whole PISA data, there are 136,344 (out of 485,490) students with complete values for this feature set.

3. Methodology

3.1. Unsupervised learning from cognitive data for label creation

We define identifying the students that are likely to succeed or fail math items of certain difficulty as a prediction problem. Our goal is to train a supervised learning algorithm that predicts success or failure from the data. However there are several problems with identifying the labels necessary for this approach. First, the plausible values cannot be used, since that would be akin to engineering an already known formula (see Section 1). Second, as discussed in Section 2, the students were administered different cognitive tests and the single items in the tests vary in their difficulty (OECD, 2014), which is why we cannot simply use the total sum of correct items for each student as their label. The raw scored cognitive data has a high percentage of missing data and no aggregated test scores and no item difficulties are available. Besides the PVs, the only available information about the actual performance of each student in the cognitive test is the fact whether he or she was administered an item and—in case the item was administered—the score the student obtained for it. The score values can be either 0 (fail), 1 or 2 (partially or fully correct).

To be able to work with the available data, we designed an algorithm to extract labels from raw data and use these labels to train a predictive model. For every different test/booklet, we summed up the total scores of the included math items. Then, we assigned each math item that was included in the test—a summary of the cluster of different items of the main tests was provided in Table 1—to a bin which we denote as *difficulty level* in such a way that each difficulty level is of same size (i.e., includes the same number of items). We chose the number of difficulty levels for our label matrix Λ to be seven, because the OECD defined seven math proficiency levels (see Figure 15.4 in the PISA 2012 technical report by the OECD (2014)). Hereby, it is assumed that all of the different booklets are consistent with regard to their average difficulty, which is supported by the fact that each test should be fair and solvable within two hours.

We created a binary label for each student and each of the seven difficulty levels, which takes value 1 if the student answered more than half of the questions in that category correctly and 0 otherwise. The labels were stored in the seven-dimensional label matrix Λ. Basically, we consider the student to be able to solve items of a certain difficulty if he or she answered the majority of the items of this difficulty bin in his/her particular test correctly. This matrix is complete, i.e. with no missing values, since each booklet contains items from each category. Depending on the target group we are interested in, we either create our label matrix Λ only for one country (for instance, for Finland the $8,829 \times 7$ matrix) or for a bigger group (for example, for all PISA countries the $485,490 \times 7$ matrix).

3.2. Supervised learning for multi-label prediction

Having the label matrix Λ fixed, we have to decide which kind of classifier should be trained for our data. Many different supervised learning algorithm have been introduced in the ML literature (Kotsiantis et al., 2007). However, the performances of different prediction models can vary depending on the data and their preprocessing. A model that performs perfectly on one dataset might perform very poorly on another dataset. Since we could not know what the best model and preprocessing for our data were, we first compared different approaches for the Finnish subset of PISA (see Section 4) before we selected the best approach to produce the final results.

In Zaki and Meira (2014), classification techniques have been categorized into probabilistic classification, decision tree classifier, linear discriminant analysis (LDA), and support vector machines (SVM). We chose at least one from each of these categories of classifiers with different objectives and compared their performances in terms of their prediction accuracy. Altogether, we compared two probabilistic classifiers (nearest neighbour and naïve bayes), one LDA, one SVM, and one decision tree based classifier (random forest). For each of the different classifiers, the Finnish subset of PISA was randomly divided, so that two thirds of the data was used for training the classifier, and one third was used for testing it.

The most important step for learning from the data is the dimension reduction in the feature space. We were looking for the minimal set of features to represent our data, since redundant or even noisy features lower the accuracy of prediction models, make them less comprehensible, and increase the computational complexity. Generally, dimension reduction methods can be divided into those techniques that extract features and those that select features (Tang et al., 2014). To get the best results, we tested with each classification algorithm two *feature extraction*—i.e., Principal

Component Analysis (PCA) and Isomap—and four *feature selection* methods—i.e., Fisher (Duda et al., 2000), Anova (Elssied et al., 2014), Gini (Hall, 1999), and MRMR (Peng et al., 2005).

3.3. Difficulty levels are predictive

Correct answers for easier questions are predictive for harder ones. With the intention to predict the performance of the students in each difficulty level as accurately as possible, we implemented an additional set of classifiers, which were the same as described above but with the difference that for each classifier, the information if the student mastered the previous difficulty level(s) was iteratively added to the original set of 53 features. That means that for predicting difficulty level λ_6 we had 54 features, for predicting λ_5, we had 55 features, and for predicting λ_1, we had 58 features. The order of the difficulty levels is $\lambda_1 < \lambda_2 \ldots < \lambda_7$, with λ_1 being the easiest and λ_7 being the most difficult one.

4. Results

We tested our algorithmic approaches by using the Finnish subset in PISA only, and then we applied the best approach first, to the Finnish (Section 4.3) and second, to the whole PISA data (Section 4.4). In Table 2, the results of the experiments with the different classifiers and dimension reduction methods are reported. As can be seen from the table, with respect to the classifier, SVM performed overall the best.

Moreover, we made the observation that the prediction accuracy was for all models the best for the highest difficulty level λ_7 and the worst for the second easiest one λ_2. The prediction accuracy for λ_1 went up again, probably because the classifiers had learned that most of the students succeed in the math items of the easiest category.

4.1. Iterative approach

To test our hypothesis that the information whether or not the student had mastered the previous difficulty level can enhance the accuracy of our classifier for the next difficulty level (see Section 3.3), we iteratively added–before predicting the next item difficulty—the previous item difficulty vector(s) as a further feature(s) to the classifiers. Naturally, testing and training data were divided according to the same indices as our original feature and label matrix. With this adjustment, the prediction accuracy improved noticeably (on average $2 - 5\%$) for difficulty level six to two for all classifiers. For difficulty level seven, the features remained the same and the accuracy of the classifier could not improve. For difficulty level one, the accuracy of the classifier actually dropped slightly. A possible explanation for that fact is, as discussed in Section 4, the general difficulty to predict the performance on the second easiest math difficulty level λ_2 correctly, as well as the observation that the prediction accuracy of the easiest difficulty level λ_1 was very high in the non-iterative approach.

4.2. Feature selection

As pointed out in Section 3.2, to avoid overfitting, we are interested in a prediction model that uses the most important features only. Therefore, we saved from all of our classifiers all features that were selected by the four feature selection algorithms in each iterative step. Then, when building the final prediction model we used for each iterative step only those features that were chosen by the different feature selection algorithms (see Section 4.3). Moreover, for training the prediction model two thirds of the data were used, and for testing it the remaining third of the data was used.

In Figure 1, the histogram of all the selected features for all iterative steps and all 53 initial features is shown. As can be seen from the histogram, the variable *Maths Self-Concept - Get Good Grades* is the most chosen feature by the feature selection algorithms, and therefore the most important variable in our math performance prediction model. Furthermore, it can be seen that, for instance, the feature *Subjective Norms - Parents Like Mathematics* is never chosen by any of the feature selection algorithms and that this feature therefore, seems to be negligible/insignificant when predicting the math performance of Finnish students.

Figure 2 also illustrates the sum of chosen features by the different feature selection algorithms. However, in this figure also the additional features λ_7-λ_2 are included. As can be seen, the information whether a student was able to master the preceding difficulty levels, are important features for the math performance prediction of the next difficulty level. It should be noted that the sums of the lasts six features cannot be fully compared, because λ_7 had the chance to be selected in all of the six last prediction models, while λ_2 could be selected only in the very last prediction models.

4.3. Results for Finland

In Table 3, the final results of the best approach for the Finnish data, i.e. the iterative SVM classifier with only the features that had been chosen at least five times (original features) or at least three times (additional λ features) by the feature selection algorithms, are reported. In each iterative step, only the features that were selected for this step were included. The table shows the accuracy, precision, recall, and f-score, which were computed on the confusion matrix of the test data.

As expected, the accuracy results are better for the higher

Table 2. Comparison of prediction accuracy (Finnish students performance in math items of different difficulty defined in label matrix Λ) with different classifiers and feature selection algorithms. The best accuracies for each level are underlined.

	Predicting success in math items of difficulty level 7						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.936816525	0.935601458	0.935601458	0.930741191	0.933171324	0.940461725	0.934386391
Naïve Bayes	0.749696233	0.933171324	0.917375456	0.764277035	0.776427704	0.940461725	0.767922236
LDA	0.919805589	0.917375456	0.899149453	0.883353584	0.878493317	0.940461725	0.876063183
SVM	0.940461725	0.939246659	0.940461725	0.940461725	0.940461725	0.940461725	0.940461725
Random Forests	0.938031592	0.940461725	0.939246659	0.933171324	0.929526124	0.940461725	0.931956258
	Predicting success in math items of difficulty level 6						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.834750911	0.825030377	0.835965978	0.82746051	0.817739976	0.838396112	0.817739976
Naïve Bayes	0.720534629	0.843256379	0.831105711	0.731470231	0.742405832	0.838396112	0.742405832
LDA	0.808019441	0.809234508	0.833535844	0.784933171	0.795868773	0.838396112	0.795868773
SVM	0.838396112	0.837181045	0.838396112	0.838396112	0.838396112	0.838396112	0.838396112
Random Forests	0.834750911	0.832041312	0.832320778	0.812879708	0.834750911	0.838396112	0.815309842
	Predicting success in math items of difficulty level 5						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.696233293	0.690157959	0.716889429	0.693803159	0.708383961	0.64763062	0.696233293
Naïve Bayes	0.662211422	0.722964763	0.705953827	0.673147023	0.670716889	0.708383961	0.67436209
LDA	0.699878493	0.688942892	0.705953827	0.690157959	0.693803159	0.708383961	0.685297691
SVM	0.722964763	0.716889429	0.710814095	0.721749696	0.722964763	0.713244228	0.719319563
Random Forests	0.722964763	0.701470231	0.714459295	0.720534629	0.704738761	0.713244228	0.722964763
	Predicting success in math items of difficulty level 4						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.614823815	0.611178615	0.575941677	0.592952612	0.599027947	0.585662211	0.605103281
Naïve Bayes	0.648845687	0.626974484	0.640340219	0.668286756	0.660996355	0.619684083	0.659781288
LDA	0.640340219	0.650060753	0.634264885	0.620899149	0.636695018	0.619684083	0.645200486
SVM	0.67800729	0.679222357	0.643985419	0.653705954	0.65127582	0.623329283	0.653705954
Random Forests	0.650060753	0.646415553	0.611178615	0.64763062	0.625759417	0.623329283	0.622114216
	Predicting success in math items of difficulty level 3						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.648845687	0.656136087	0.602673147	0.635479951	0.657351154	0.652490887	0.669501823
Naïve Bayes	0.64763062	0.652490887	0.671931956	0.64763062	0.645200486	0.652490887	0.650060753
LDA	0.637910085	0.631834751	0.662211422	0.637910085	0.643985419	0.659781288	0.65127582
SVM	0.675577157	0.668286756	0.662211422	0.665856622	0.65127582	0.64763062	0.667071689
Random Forests	0.667071689	0.648845687	0.611178615	0.67436209	0.62818955	0.641555286	0.643985419
	Predicting success in math items of difficulty level 2						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.573511543	0.583232078	0.539489672	0.543134872	0.539489672	0.546780073	0.546780073
Naïve Bayes	0.571081409	0.582017011	0.622114216	0.569866343	0.579586877	0.602673147	0.583232078
LDA	0.577156744	0.580801944	0.602673147	0.59781288	0.589307412	0.607533414	0.57472661
SVM	0.59781288	0.59781288	0.605103281	0.596597813	0.599027947	0.605103281	0.599027947
Random Forests	0.572296476	0.599027947	0.545565006	0.591737546	0.571081409	0.603888214	0.567436209
	Predicting success in math items of difficulty level 1						
	Full	PCA	Isomap	ANOVA	Fisher	MRMR	Gini
Nearest Neighbors	0.733900365	0.738760632	0.720534629	0.732685298	0.713244228	0.769137303	0.733900365
Naïve Bayes	0.606318348	0.753341434	0.769137303	0.617253949	0.616038882	0.769137303	0.618469016
LDA	0.714459295	0.708383961	0.741567436	0.716889429	0.733900365	0.769137303	0.730255164
SVM	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303	0.769137303
Random Forests	0.769137303	0.763061968	0.737545565	0.760631835	0.732685298	0.759416768	0.739975699

difficulty levels (because most students will fail this level) and the lower difficulty levels (because most students will master this level) than for the middle difficulty levels. On the other hand, the precision increased monotonically from the most difficult to the easiest question difficulty level. This was most probably the case, because the classifier had learned that most students fail items of the highest difficulty and hence, simply returned 0 for the majority of the test instances. Since accuracy is not the best measure of performance we focus on the precision for the rest of the discussion.

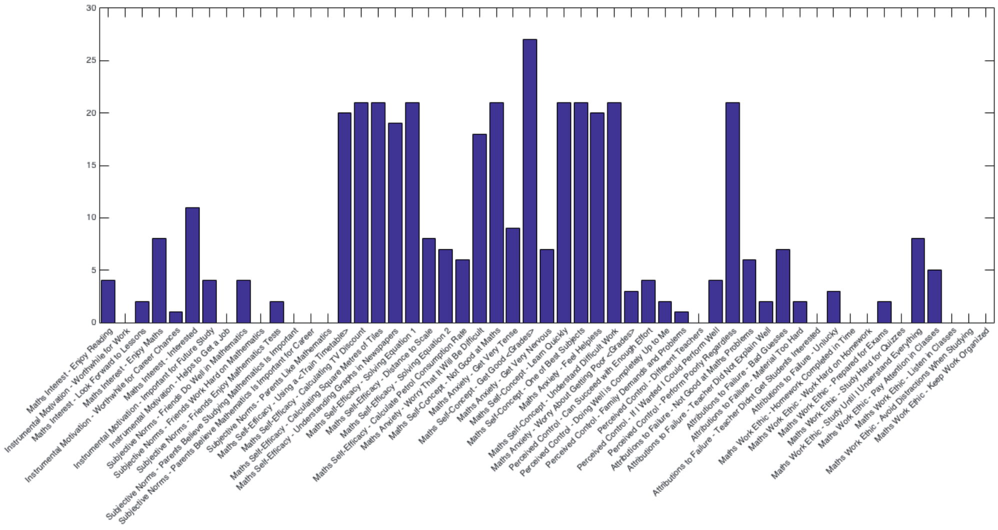


Figure 1. Frequency of selected features of the 53 initial features by the four feature selection algorithms for the Finnish student data. The higher the bar of a feature, the more often this feature was selected, and the more important this feature is for the prediction model.

Table 3. Results of iteratively predicting success in math items of the different difficulty levels for Finnish students with SVM and—for each difficulty level—only the most selected features by the four feature selection algorithms.

Difficulty	**Accuracy**	**Precision**	**Recall**	**F-score**
Level 7	0.9579	0.0312	1.0000	0.0606
Level 6	0.8555	0.1385	0.5294	0.2195
Level 5	0.7427	0.3309	0.6866	0.4466
Level 4	0.7843	0.4029	0.6975	0.5108
Level 3	0.7096	0.5496	0.7791	0.6445
Level 2	0.6757	0.6530	0.7095	0.6801
Level 1	0.7630	0.9493	0.7833	0.8583

4.4. Results for all countries participating in PISA

Table 4 shows the prediction results for all PISA countries (i.e. the 136344×53 feature matrix of all students that had complete values for all 53 features from the background questionnaire and the corresponding 136344×7 label matrix for the same students). However, it should be noticed that the same settings as for Finland were used, that is the classification algorithm and the selected features that were optimized for the Finnish data. For difficulty levels λ_6 and λ_5 the prediction accuracies are actually higher than for the Finnish data. However, this is most likely the case because most of the world's students are not able to solve items of this difficulty level. This assumption is supported by the very low precision values. Moreover, we see again the worst result for predicting λ_2, where the prediction accuracy is only slightly better than guessing.

Table 4. Results of iteratively predicting success in math items of the different difficulty levels for students from all in PISA participating countries with SVM and—for each difficulty level—only the most selected features by the four feature selection algorithms.

Difficulty	**Accuracy**	**Precision**	**Recall**	**F-score**
Level 7	0.9524	0.0003	0.0714	0.0006
Level 6	0.8872	0.0027	0.2414	0.0054
Level 5	0.7723	0.0133	0.3118	0.0255
Level 4	0.6156	0.1627	0.5016	0.2457
Level 3	0.5817	0.7934	0.5918	0.6779
Level 2	0.5350	0.5629	0.5404	0.5514
Level 1	0.6539	0.9591	0.6656	0.7859

5. Discussion and future work

PISA data—as well as LSA data generally—provide an interesting source for educational ML and DM applications, because they are of high quality, internationally comparable, and publicly available. However, the challenges of working with these data are the high sparsity of the raw data and the lack of any readily available and comparable cognitive test results of the students.

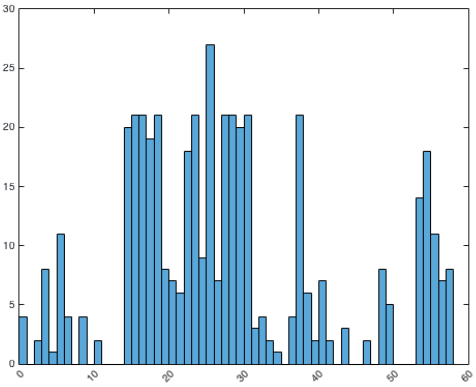


Figure 2. Histogram of selected features of the 53 initial features plus the 6 additional ones for the iterative steps by the four feature selection algorithms for the Finnish student data.

In this paper, we have presented an approach to prepare LSA data for supervised ML approaches. In addition, initial results of using our approach for predicting success in math items of various difficulty, have been presented. Hereby, we have tested different classification and dimension reduction algorithm for the Finnish data, and then applied the best classifier with only the selected features of different feature selection algorithm for the Finnish and for the whole PISA data. The prediction accuracy was further improved by adding for each succeeding difficulty level the information whether the student mastered the preceding difficulty level(s). An analysis of the chosen features by the feature selection algorithm enabled a predictive power ranking of the questions asked in the background questionnaire that actually explained the students' math performance.

The results presented in this paper are only preliminary and we intend to extend and improve our experiments and study in various directions. First of all, the results that were presented here are based on the fully available raw data only. We intend to perform similar experiments for the whole contextual data by first imputing the missing values.

We also intend to compare our approach to the Rasch model and plausible value approach currently used in most LSAs, which has evolved from the psychometric literature. It has been argued that one of the weaknesses of the Rasch model is the fact that all students with the same raw score (i.e., number of correctly solved tasks) obtain the same ability estimate (Embretson & Reise, 2013). It would be interesting to compare this to our approach, where the difficulty level of the solved items is taken into account. As discussed by Baker and Yacef (2010), comparing and integrating machine learning techniques to the ones from the psychometrics literature, is one of the most distinguishing features that separates the educational ML/DM discipline from the traditional ML/DM research area.

References

Baker, R. and Yacef, K. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1, 2010.

Duda, Richard O., Hart, Peter E., and Stork, David G. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000. ISBN 0471056693.

Elssied, Nadir Omer Fadl, Ibrahim, Othman, and Osman, Ahmed Hamza. A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3):625–638, 2014.

Embretson, Susan E and Reise, Steven P. *Item Response Theory*. Psychology Press, 2013.

Hall, Mark A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

Kärkkäinen, T. and Saarela, M. Robust Principal Component Analysis of Data with Missing Values. In *Machine Learning and Data Mining in Pattern Recognition*, pp. 140–159. Springer, 2015. ISBN 978-3-319-21023-0. doi: 10.1007/978-3-319-21024-7_10.

Kotsiantis, Sotiris B, Zaharakis, I, and Pintelas, P. *Supervised machine learning: A review of classification techniques*. OS Press, 2007.

Marsman, Maarten. *Plausible Values in Statistical Inference*. Universiteit Twente, 2014.

Monseur, Christian and Adams, Ray. Plausible Values: How to Deal with Their Limitations. *Journal of applied measurement*, 10(3):320–334, 2008.

Musik, Alexander. Philologenverband bezeichnet Pisa-Studie als Geldverschwendung. `http://www.deutschlandfunk.de/bildungsforschung-in-der-kritik-philologenverband.680.de.html?dram:article_id=347675`, 2016.

OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009. ISBN 9789264056268.

OECD. *PISA 2009 Technical Report*. OECD Publishing, 2012.

OECD. PISA 2012 Technical Report, 2014.

Peng, Hanchuan, Long, Fuhui, and Ding, Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Rasch, Georg. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. 1960.

Rutkowski, Leslie, Gonzalez, Eugenio, Joncas, Marc, and von Davier, Matthias. International Large-Scale Assessment Data Issues in Secondary Analysis and Reporting. *Educational Researcher*, 39(2):142–151, 2010.

Saarela, M. and Kärkkäinen, T. Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 60–68, 2014.

Saarela, M. and Kärkkäinen, T. Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 156–163, 2015a.

Saarela, M. and Kärkkäinen, T. Weighted clustering of sparse educational data. *In 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 337–342, 2015b.

Tang, Jiliang, Alelyani, Salem, and Liu, Huan. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, pp. 37, 2014.

Tarpey, Thaddeus and Petkova, Eva. Latent regression analysis. *Statistical modelling*, 10(2):133–158, 2010.

Wu, M and Adams, RJ. Plausible values: Why they are important. In *11th International Objective Measurement Workshop, New Orleans*, 2002.

Zaki, Mohammed J and Meira Jr, Wagner. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

PIX

SUPPORTING INSTITUTIONAL AWARENESS AND ACADEMIC ADVISING USING CLUSTERED STUDY PROFILES

by

Mariia Gavriushenko, Mirka Saarela, Tommi Kärkkäinen 2017

Proc. of the 9th International Conference on Computer Supported Education,
Volume 1, pp. 35–46

Supporting Institutional Awareness and Academic Advising Using Clustered Study Profiles

Mariia Gavriushenko, Mirka Saarela and Tommi Kärkkäinen
Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
magavriu@student.jyu.fi, mirka.saarela@jyu.fi, tommi.karkkainen@jyu.fi

Keywords: Academic Advising, Learning Analytics, Robust Clustering.

Abstract: The purpose of academic advising is to help students with developing educational plans that support their academic career and personal goals, and to provide information and guidance on studies. Planning and management of the students' study path is the main joint activity in advising. Based on a study log of passed courses, we propose to use robust, prototype-based clustering to identify a set of actual study path profiles. Such profiles identify groups of students with similar progress of studies, whose analysis and interpretation can be used for better institutional awareness and to support evidence-based academic advising. A model of automated academic advising system utilizing the possibility to determine the study profiles is proposed.

1 INTRODUCTION

The credit-based system is used to characterize the requirements and progress of a student in many learning environments. Availability of personalized support is very important constituent of a learner's success (Nguyen et al., 2008). Academic advising (AA) is an iterative collaboration process between student, academic adviser, and academic institution, to tackle the student retention. Advisers provide versatile assistance to the students during their studies, making the educational experience relevant and supported.

AA activity has a long history dating back to 1870s (Tuttle, 2000). Advising starts, when a student becomes enrolled in higher education, and finishes when the degree has been completed. The purpose of academic advising is to ensure that the students carry out the required studies to graduate. The central activity for this purpose is to support study planning, especially at the beginning of the academic life. Depending on the organizational culture, especially the stability or dynamicity of the course schedule, an academic adviser either needs to ensure that predefined study plan is being followed or that a student knows all the relevant study possibilities. Email, social media, web and wikipages etc. provide means to share the necessary information with the students. However, typically face-to-face discussions take place either regularly (e.g., at the beginning of a semester or academic year) or by students' or advisers' request. For more personalized support, an adviser should know when a student is in need of a study advice discussion and what is the precise status of the studies.

Preliminary recommendations to do certain, especially compulsory, courses to proceed normally with the major subject studies are provided by departments and advisers. More precisely, for example at the University of Jyväskylä (JYU), all students are required to prepare an electronic personal study plan with the academic adviser from the home department. Advising is organized according to the *satellite model* referring to the distributed responsibility of academic units (Tuttle, 2000). The *engagement model* between advisee and adviser characterize the principal spirit of counselling (Feghali et al., 2011).

In JYU, the study plan and the completed studies create the starting point for a study plan assessment discussion. However, especially in computer science, the actual number of studies that have been made during an academic year are typically less than recommended (Saarela and Kärkkäinen, 2015a). Hence, it is very common that the actual study path deviates from the recommendations and plans, and in such a case an advising intervention is needed. But how does an individual student and especially an individual study adviser know, what is the relation between students' realized study path and that of the peer students? Thus, could we, instead of comparing against the predefined plans, advise students based on evidence from the actual study paths of other similar students?

Therefore, the purpose of this article is to propose using a learning analytics method (Chatti et al., 2012), more precisely robust clustering (Äyrämö, 2006; Saarela and Kärkkäinen, 2015a), to create groups of actual profiles of students concerning their studies. Such profiles summarize the different typical accumulations of completed studies, increasing the general awareness of the common study flows. One can explicitly link an individual student to student peers with similar study path profile in the same institutional environment. This allows an adviser to plan a possible intervention adaptively for a larger pool of students instead of following each one individually.

Creating general study profiles of students can help departments in their assessment and planning of when (and how) they provide the courses, especially the compulsory ones (Saarela and Kärkkäinen, 2015a). By automating general student profiling it is possible to provide essential support for adaptive on-line advising along the lines suggested, e.g., in (Nguyen et al., 2008; Henderson and Goodridge, 2015). Individual student's perspective, from self-regulated learning and study planning point of view without academic advising interface, was thoroughly addressed in (Auvinen, 2015) (see also (Auvinen et al., 2014)).

The contents of this article is as follows: after the introduction, we provide background on academic advising and personalized student support in Section 2. Then, in Section 3, we describe the data and the robust clustering method, and introduce three cases to construct student profiles to support academic advising. In Section 4 features and a model of an Academic Adviser system with automated mechanisms in it are proposed. The work is concluded in Section 5.

2 BACKGROUND

Academic advising is a collaborative process in which adviser and advisee enter a dynamic relationship where adviser helps advisee to enhance the learning experience by helping in making academic decisions (Henderson and Goodridge, 2015). The decision support could be made by analysis of student's records, as well as some external factors like interests, goals, academic capabilities, schedules etc. (Noaman and Ahmed, 2015). Developmental Advising means helping students to define and explore academic and career goals and pathways, as well as to develop problem-solving and decision-making skills; Prescriptive Advising, which is the more traditional advising model, is mainly concentrated on providing the information to the students according to their academic program, progress, academic policies, course selection, etc.; Intrusive Advising refers to contacting with student in critical periods like first year of study before the declaring major, graduation period, or when students are at-risk or they are high-achieving students (Noaman and Ahmed, 2015).

Next we briefly summarize a pool of directly AA related work that was identified through a non-systematic search. Our main concern here is to illustrate the strong link between the needs and practices of AA and the general utility of student profiling.

2.1 On Academic Advising

Student voices on AA were raised in only some articles. El-Ansiri et al. (Al-Ansari et al., 2015) used questionnaire to study student satisfaction and support-seeking patterns among dental students in Saudi-Arabia. Very low (only 7.6%) primary utility help rate of advisers in the academic matters was encountered. Even if the advisers were available when needed, they were not able to provide the most relevant information, e.g., on important dates and courses. Hence, up-to-date course and timetable information seems to be a prerequisite for AA, which is handled by the in-house developed, integrated study information system *Korppi*[1].

The pedagogical side of AA was also focused rarely. In the work (Drozd, 2010) author studied academic advisers through the lens of transformational leadership, i.e. how advisers can create a connection to students that positively influence their study paths (by increasing and inspiring study motivation and engagement/commitment in studies through individual and intellectual consideration). A questionnaire for undergraduate students strengthened the importance of transformational leadership activities in adviser-student communication and collaboration, independently from the student's characteristics. The lack of time for individual counselling efforts that was visible in most of the reviewed articles here was not emphasized in (Drozd, 2010). Dougherty (Dougherty, 2007) studied academic advisers from those students' perspective who are doing very well in their studies. These students are called high-achieving students. Authors address the need for the investigation of unique characteristics of these students.

Technical support for AA has been considered in many articles. The availability of extensive information on courses to support automatization of AA was emphasized in (Biletskiy et al., 2009). The authors proposed course outline data extractor application, which helps in recognizing similar or comparable courses between different institutions, also help-

[1] `https://www.jyu.fi/itp/en/korppi-guide`

ing both students and academic advisers to keep track of the variety of topics that i) have been covered in the completed studies, ii) should be covered to complete minor or major subject modules or the actual degree. The authors in (Nguyen et al., 2008) proposed an integrated knowledge-based framework based on semantic technology that supports computer-based (automatic) e-Advising on the suitable courses for the students. Naturally individual learning history data provide the starting point for the system and, for this purpose, the authors implemented and tested a data integration tool.

The high workload of academic advisers, especially due to individual but many times recurrent handling of basic issues with multiple students in a hurry, was addressed in (Henderson and Goodridge, 2015), with the proposition of an intelligent, semantic, web-based application to assist decision making and automatization of repetitive counselling tasks. Core of the system consisted of rule-based inference engine, which mapped student profile with the study program profile and organizational rules, to provide automatic suggestions on the courses to be enrolled in the upcoming semester. In the preliminary evaluation, a positive feedback of the system was obtained, although the main limitation of suitability to only study programs which follow a clear, predefined study path of courses, was recognized. With very similar aims and functionality, another web-based on-line adviser was described in (Feghali et al., 2011). This system was also evaluated positively when compared to the current advising system. The authors emphasized that such a tool only supports and does not replace a human academic adviser.

Conversational, fully autonomous agent supporting AA dialogs using natural language processing (NLP) were suggested in (Latorre-Navarro and Harris, 2015). The proposed system contained an extensible knowledge base of information and rules on academic programs and policies, course schedules, and a general FAQ. NLP performance of the proposed system was evaluated positively. Also, the similar multi-agent approach was suggested in (Wen and McGreal, 2015) for AA. This approach helps tackling a dynamic and complex individualized study planning and scheduling problem. As well as in (Al-Sarem, 2015) was proposed a decision tree model for AA affairs based on the algorithm C4.5. The output is evaluated based on Kappa measure and ROC area. The main conclusion was made that the difference between the registered and gained credit hours by a student was the main attribute that academic advisers can rely on (Al-Ansari et al., 2015).

As can be concluded, earlier studies have mostly concentrated on research prototypes which focus only on few main components or tool support for existing learning management systems. Taking into account that user modeling is one of the key factors for including personalization into the learning system, many researchers used ontologies for learners' models, because ontologies have many advantages for creation of user models (Idris et al., 2009; Chen, 2009a; Leung et al., 2010; Nguyen et al., 2008; Biletskiy et al., 2009; Henderson and Goodridge, 2015).

Data-mining techniques have also been applied to the learning environments in order to track users' activities, extract their behavior profiles and patterns, and analyze the data for future improvement of the learning results, as well as for identifying types of learners (Minaei-Bidgoli, 2004). Mostly, for developing personalized learning plan, researchers used decision tree search, heuristic algorithms, genetic algorithms, item response theory and association rules. Also, many studies used semantic web technologies, neural networks and multi-agent approach. Most of the previous studies on personalized learning path generation schemes have mainly focused on guiding the students to learn in the digital world; i.e. each learning path represents a set of digitalized learning objects that are linked together based on some rules or constraints (Liu et al., 2008). While determining such digitalized learning paths, the learning achievements, on-line behaviors or personalized features (such as learning style) of individual students are usually taken into consideration (Schiaffino et al., 2008; Chen, 2008; Chen et al., 2008; Chen, 2009b; Chen et al., 2005).

2.2 On Personalization of Student Support

In general, many researchers have paid attention to developing *e-learning systems with personalization*, and the most common aspect in these system is the *creation of the personalized learning path* for each individual student or group of students. Most of personalized systems consider learner preferences, interests and browsing behaviors, because it will help to provide personalized curriculum sequencing service (Huang et al., 2007). In the study (Chen et al., 2005) authors proposed a personalized e-learning system which is based on Item Response Theory (PEL-IRT). This system is considering course material difficulty and learner ability, to provide individual learning path for learners. Learner's ability estimation was based on an explicit learner's feedback (the answers of learners to the assigned questionnaires). The system appeared mostly like a recommendation system of the courses for the learners. Authors in (Huang et al., 2007) proposed a genetic-based curriculum sequencing approach and used case-based reasoning to develop a summative assessment. The empirical part indicated that the proposed approach can generate appropriate course materials for learners taking into account their individual requirements. Later, in (Chen, 2009a), the authors developed a personalized web-based learning system grounded on curriculum sequencing based on a generated ontology-based concept map, which was constructed by the pre-test result of the learners. Optimization problem for modeling criteria and objectives for automatic determination of personalized context-aware ubiquitous learning path was suggested in (Hwang et al., 2010). This learning model not only supports learners with alternative ways to solve problems in real-world situations, but also proposes more active interaction with the learners. Authors in (Werghi and Kamoun, 2009) proposed Decision Support System for student advising based on decision tree for an automated program planning and scheduling. The proposed approach takes into account prerequisite rules, the minimum time (minimum number of terms), and the academic recommendations. The adaptive course sequencing for personalization of learning objectives was suggested in (Idris et al., 2009) using neural networks, self organizing maps and the back-propagation algorithm.

A very closely related work to ours was reported in (Sandvig and Burke, 2005). Authors proposed a case-based reasoning paradigm which is based on the assumption that similar students will have similar course histories. The system used the experience and history of graduated students in order to propose potential courses for the students. Unfortunately, this approach required matching between students' histories. Also, similar case-based reasoning was used by (Mostafa et al., 2014) for developing a recommendation system for a suitable major to students based on comparison of the student information and similar historical cases.

As reviewed, many suggestions for intelligent software and information system support of AA have been given. Many studies describe the creation of intelligent learning systems that can make a curriculum sequencing more flexible for providing students with personalized and adaptive study support services (Fung and Yeung, 2000; Lee, 2001; Brusilovsky, 1998; Lee, 2001; Papanikolaou et al., 2002; Tang and McCalla, 2005). Universities are more and more looking into developing self-service systems with intelligent agents as an addition or replacement for the labor-intensive services like academic advising. For example, The Open University of Hong Kong has developed an intelligent on-line system that instantly responds to enquiries about career development, learning modes, program/course choices, study plans, and graduation checks (Leung et al., 2010).

However, the institutional starting point concerning available digital information, especially for the web-based systems that have been proposed, seems to vary a lot. Some systems start and focus on providing easy access to course and degree requirements information whose availability is to be assured first. On the other hand, we might start from the situation where we can readily access most of the relevant data: i) course information with basic contents, learning goals, assessment methods, acceptance criteria, schedule and location, teachers and lecturers etc.; ii) individual, anonymous study records on passed courses and completed studies. (Note that reliable information on student admission is currently not directly available in the organization under consideration).

3 CREATION OF STUDY PATH PROFILES USING ROBUST CLUSTERING

3.1 Data

To illustrate the proposed approach, we utilized real study records of the Bachelor (BSc) and Master (MSc) students majoring in Mathematical Information Technology (which is comparable to a major in Computer Science at other universities) at the University of Jyväskylä (JYU/MIT). IT administration at the University has recently created a data warehouse of passed courses by all the students, which can be utilized by the departments. On the other hand, the electronic study plan system does not provide direct interface for larger student groups, so both from accessibility and evidence-basedness points of view, we focus on analyzing the real study log of the passed courses. The log was anonymized, keeping student IDs as keys, covering the four calendar years 2012–2015. Note that students can start their studies in the beginning of September (autumn term) or January (spring term). Hence, the original study registry log included a heterogeneous set of BSc and MSc students who had started their studies either before 2012 or in the beginning of spring or autumn terms during 2012 – 2015.

The whole study log contained 15370 passed courses by 1163 different students on 1176 different course IDs. There were 942 male students (81%),

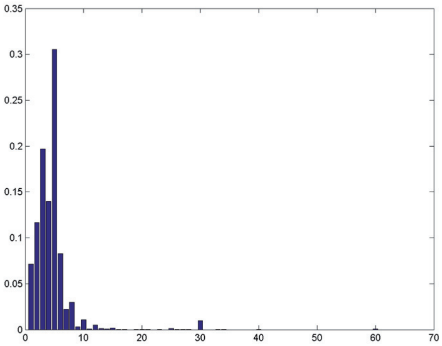


Figure 1: Probability of size of a course.

with mean amount of studies made 59.9 ECTS. Only 221 female students (19%) were identified, with mean amount of studies made 57.0 ECTS. Hence, most of the students in the log were either in the beginning of their studies or progressing very passively and slowly.

Figure 1 shows the discrete density distribution of the size of the passed courses. According to the figure, 5 ECTS and 3 ECTS are the two most common sizes of the courses, the former covering around 30% of the studies. Moreover, there are a lot of small courses (1 – 6 ECTS) with the exception of the MSc thesis, 30 ECTS. Teaching in JYU is organized for four periods during one academic year (plus the summer semester) in such a way that a course of ca. 5 ECTS can fit to one period. We conclude that because the passed courses represent both major and minor subject studies, division of the overall learning objects as courses is not optimal. This observation is the first example on how summarization of study log data provides visibility and feedback to the organization. During the course of writing this article, we also found out that the instructions of JYU for preparing the next curriculum for 2017–2019 include strong recommendation to decrease the number of courses with only a few credits.

Next we aggregated how many credits per semester each student had made. Similarly to (Saarela and Kärkkäinen, 2015a), each calendar year was divided into two semesters: the spring term (from January to June) and the autumn term (from July to December). However, since usually only a few courses are completed during the summer (this is illustrated in (Saarela and Kärkkäinen, 2015a)), it was reasonable to divide the calendar year into only two parts for further analysis.

In what follows, we profile, analyse and compare two students cohorts: those who started their studies in the beginning of the autumn term 2012 (A2012) or 2013 (A2013). Hence, for A2012 we end up with 8 and for A2013 with 6 integer variables representing the aggregated amount of credits on half-a-year scale. Since the students have progressed in their studies very differently and many of them have not been active during all the semesters of interest, both of the data sets are very sparse containing a lot of missing values (Saarela and Kärkkäinen, 2015a). This is the key property that is taken into account in the profiling approach that is described next.

3.2 Robust Clustering Method

As already explained, our goal is to assist the academic advisers by recommending suitable courses for students based on passed courses of (possibly more advanced) students with similar study path. For this, we need to identify general profiles of similar students and this is, precisely, the purpose of clustering. Partitional (or representative-based (Zaki and Meira Jr, 2014)) clustering seems to be the right family of clustering methods to choose from because it assigns each observation to exactly one cluster, which is represented by its most characteristic point, the cluster centroid, which represents the common profile. Within a cluster, distances of observations to the prototype determine the most typical or representative members of a cluster. Thus, instead of following many different student profiles, the academic adviser can just follow the most common profiles to get an overview of the whole cohort.

Generally, partitional based clustering algorithms consist of an initialization step, in which the initial centroids of each cluster are generated, and iterations of two steps where (i) each observation is assigned to its closest centroid, and (ii) the centroid of each cluster is recomputed by utilizing all observations assigned to it. The algorithm stops when the centroids remain the same over two iterations. The most popular and most applied partitional clustering algorithm is the *k-means* (Jain, 2010), also in learning analytics studies (Saarela and Kärkkäinen, 2017). This algorithm works very well for full and approximately normally distributed data since the sample mean is the most efficient estimator for samples that are drawn from the normal distribution. However, the sample mean is highly sensible to all kinds of outliers (Huber, 2011) as well as missing values, which can be characterized as special types of outliers. Also for a nonsymmetric (skewed) distribution, the sample mean is not necessarily the most efficient estimator and other location estimates might be preferable (Sprent and Smeeton, 2016). Moreover, as explained in (Saarela and Kärkkäinen, 2015a), the quantization error for the integer-type variables like here has uniform not gaussian distribution.

The spatial/geometric median is a robust nonparametric location estimate, which remains reliable even if half of the data is contaminated (Sprent and Smeeton, 2016). Mathematically, the spatial median is the Weber point that minimizes the (nonsquared) sum of the Euclidean distances to a group of given points $\{\mathbf{x}_i\}, i = 1, 2, \ldots n$:

$$\arg\min_{\mathbf{c}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{c}\|.$$

Although the basic concept is easily understood and has been extensively discussed in the literature (albeit under various names, see (Drezner and Hamacher, 2001)), its computation is known to be difficult.

In (Äyrämö, 2006), the difficulty of computing the spatial median during partitional clustering was solved with the SOR (Sequential Overrelaxation) algorithm (see (Äyrämö, 2006) for details). Moreover, in the implementation of the resulting *k-spatial-medians* clustering algorithm, only the available (i.e. not-missing) data is taken into account when the centroid is recomputed.

To sum up, all of these above discussed properties – most importantly, the robustness to missing data and the fact that every cluster is represented by a centroid – make the *k-spatial-medians* clustering very suitable for creating student's general study profiles. The fact that such a clustering approach works very well for sparse educational data has been previously shown in (Saarela and Kärkkäinen, 2014; Saarela and Kärkkäinen, 2015b; Saarela and Kärkkäinen, 2015). The initialization of the robust clustering method was realized similarly as in (Saarela and Kärkkäinen, 2015): We started with multiple repetitions of *k-means* for the complete data – without missing values – and then, applied *k-spatial-medians* to the best of those results.

3.3 Clustered Student Profiles

Similarly to the earlier work in (Saarela and Kärkkäinen, 2014; Saarela and Kärkkäinen, 2015b; Saarela and Kärkkäinen, 2015a; Wallden, 2016), we apply four different internal cluster validation indices to determine the number of clusters: Knee Point (KP) of the clustering error, Ray-Turi (RT), Davies-Bouldin (DB), and Davies-Bouldin∗ (DB∗). All the computations here were carried out in the Matlab-environment, using own implementations of all the algorithms.

From the two student groups A2012 and A2013, we include in clustering only still active students, i.e. those who have made credits during the autumn term 2015 (the last one analyzed). Furthermore, we restrict ourselves to those students for whom over half of the variables are available (Sprent and Smeeton, 2016). This means that the 47 analyzed students in A2012 have made studies during at least four out of the seven possible semesters (including the last one) and the 76 students in A2013 at least in three out of the five semesters. Because of the anonymity, we obtained further assistance in relation to the metadata and interpretation of the clusters from the Study Amanuensis of the Department (Study Amanuensis, 2016).

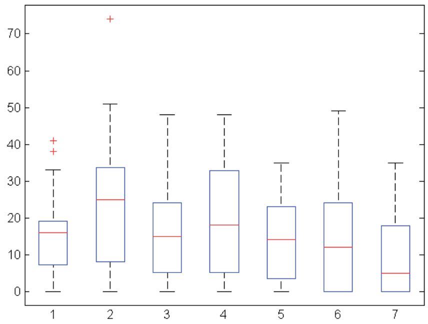


Figure 2: Boxplot for A2012.

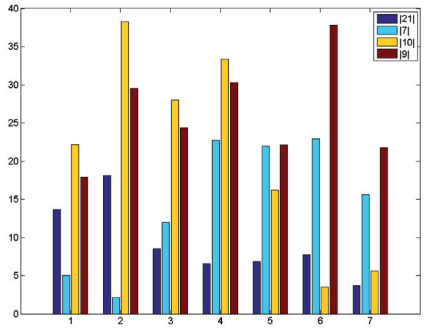


Figure 3: Credit accumulation prototypes for A2012.

A2012

The boxplot in Figure 2 shows the large variability in the study accumulations both within semesters and between semesters. We see the larger accumulations in the spring terms during the first two years, and a slightly decreasing overall trend after that. There are always exceptional students who have made much more studies than their peers.

KP, DP, and DP∗ indicated four clusters and RT had also local minimum there, so we choose to analyze four different general study progress profiles. The profiles for A2012 are depicted in Figure 3, where the size of the cluster is given in the top-right corner. The profiles are sorted in the ascending order with respect to the total number of credits.

The main group of 21 students in the first cluster illustrate a potential start of the studies in the first year, with strong passivation after that. They have obtained prototypically 65 ECTS until the end of 2015. Based on (Study Amanuensis, 2016), by a closer look on the 8 students from the cluster closest to the centroid, these are all older BSc and MSc male students (born before 1990). They are either distant students studing while working or have completely chosen to change their orientation from an earlier occupation and already finished degree. The difficulties in studies and reasons of such a behavior, for a similar adult student profile, were thoroughly discussed in the earlier work (Kaihlavirta et al., 2015) from the same context (department) than here.

The second group of only 7 students, who generally obtained 103 ECTS, shows opposite behavior: very slow start in the first year, activating to an appropriate level then. Three most characteristic students here were young males, who were involved in the military service during the first study year. This complete explains the observed behavior.

The third group of 10 students, who generally obtained 147 ECTS, did their studies very actively for the first 4–5 semesters. Analysis of the three most characteristics students revealed two young and one older male students who either took job or became active in student organizations during the third year of the studies.

The fourth profile with 9 students, altogether 184 ECTS in general, illustrates that a good start on the study activity carries over the semesters. Three mostly characteristics students were again all males, one MSc student and two BSc students. Note that similar finding on the importance of active start in an individual course level was given in (Saarela and Kärkkäinen, 2015a).

Students who are mostly in need of academic advising are the ones in the first cluster. They can be identified either in the beginning of their studies or after the second semester, because even if still making studies, their accumulation is much less than in the third and fourth cluster. Their characterization also suggests the department to rethink the study entrance criteria.

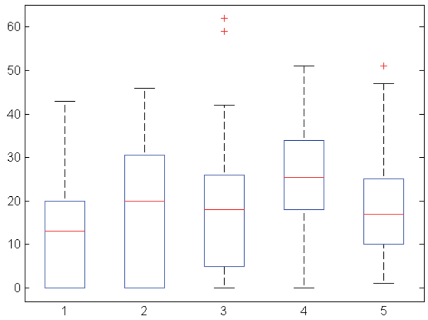


Figure 4: Boxplot for A2013.

A2013

For A2013 all cluster indices suggested three profiles, which are illustrated in Figure 5. This and the fact that there are now one profile less than in A2012 suggests more stable organization of the curriculum. Also the boxplot in Figure 4 supports such finding, especially showing smaller variability in the obtained credits between the autumn and spring terms compared to A2012 in Figure 2.

Student group in the need of intrusive academic advising consists of those 23 students with smallest accumulation of credits. These students start and continue very slowly in their studies, although the level of activity was increasing in the fourth and fifth semesters. Their general ECTS accumulation after five semesters was 44 ECTS. Analysis of the five most representative students revealed two older male students (birth year before 1990), two males with indications of military service, and a female student. According to (Study Amanuensis, 2016), especially the younger students showed signs of low self-regulation during academic advising sessions.

The second profile of 17 students, completing typically 80 ECTS, showed similar behavior to the second profile in A2012: the minimal first year is raised to a good level of study activity later. A closer look on the five most representative students showed young, two female and three male students. Four of these had identified themselves as a non-active student during the first study year, again mostly due to the military service of the young male students.

The third profile of 36 most active students, accomplishing 123 ECTS typically, showed similar overall behavior than the fourth profile in A2012. The first semester is slightly smaller but then the study path proceeds in the desired way. Recapitulation of

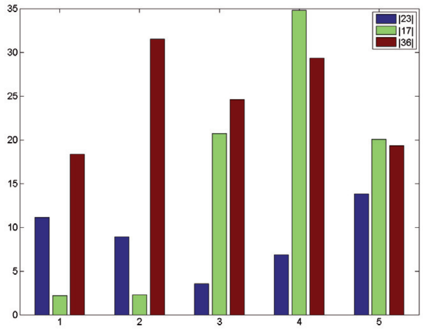


Figure 5: Credit accumulation prototypes for A2013.

the meta data of five most representative students showed five male students, of whom three were oriented towards game programming and development - the most recent study line of the department.

We note that even if the boxplot in Figure 4 indicated more stable study path with respect to autumn and spring semesters, the two profiles of truly active students still illustrated larger study accumulations in the spring than in the autumn. These findings are, however, mostly explained by the longer calender time for the two periods in the spring term compared to the autumn term – a general peculiarity of the Finnish higher education system.

The similarities and differences between the two sets of profiles just discussed emphasize the importance of the use of evidence-based information in academic advising. On one hand, there are repetitive profiles of students proceeding in their studies well or slowly. The latter ones needs to be detected and supported in an intrusive manner in academic advising. The home department responsible for major subject studies and the other departments providing minor subject studies should be informed about the found hindrances of the study paths. In the case analyzed here, there is a clear change of study accumulation profiles from A2012 to A2013, which suggests that the organization of courses, the capabilities of students, and/or their support through academic advising have improved in the educational organization under study.

4 PROPOSITION OF A SYSTEM MODEL

As shown, it is important to follow the actual progress of the students in their studies. There might be no need for an advising intervention, but if so, one should automatically notify the students and the study counselling on the deviations in the study path. The problems of not passing courses and not following suggested study plans usually also call for organizational considerations whether learner ability and the difficulty level of the recommended curriculum are matched to each other properly (Huang et al., 2007).

4.1 Proposed System Architecture

This subsection describes the novel system architecture for AA and automatic feedback based on recognized student group profiles which are obtained by using clustering. We also present an overview of the AA process as both manual and automated process.

The architecture of the proposed system's model for the AA is presented in Figure 6. The system has two main databases: learner profile database and curriculum database. Learner profile database stores learner's data about studies, assessment results, timetables of completed studies, etc. Curriculum database stores information about compulsory courses, other courses, timetables, etc. The academic advising system's part consists of several blocks like linking individual students to their peers with similar study path profile together with the recommendation block and planning block.

Based on the system architecture, the details of system's main functionality read as follows:

1. Collection of learner's personal information.
2. Collection of information about the courses and completed studies.
3. Creation of study progress profiles along the lines of Section 3.
4. Linking the individual student to student peers with similar study profile in the institutional environment.
5. Student's progress check. If student is linked to a profile requiring intrusive advising, inform the ad-

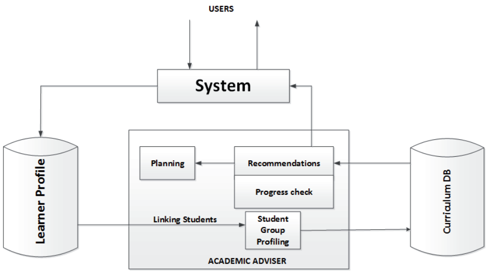


Figure 6: Architecture of the Academic Adviser.

viser and the student by providing the interpreted study profile to support the communication and problem solving.

6. Modification of the study plan on recommended courses and their timetables by taking into account the evidence related to the identified study profile.
7. Planning and realization of an intrusive profile intervention adaptively for a larger pool of students.

Data collection related to the system is, naturally, all the time active. The evidence-based study profiles can and should be recomputed on regular basis. A natural suggestion would be to do this after the studies made during the previous semester have been stored and become available for clustering.

4.2 Automated Academic Advising process

The automated process of Academic Advising, related to the system's architecture and main functionality as described above, is presented in Figure 7. The given proposition allows manual control of the continuous advising activity for every learner individually, or the more automated process where the role of the advisers is shifted to the higher level of abstraction. The difference on the level of learner's life cycle between these two use cases is depicted in Figure 8. The automated process is highlighted with the red color in the figure. In the automated scenario, the responsible persons of the study organization only provide policies, planning and regulations. This can reduce the responsibility for the daily routine work and could help to provide recommendations for a larger pool of students rather than for the each individual learner.

The work-flow related to Figure 7 reads as follows: The learner is choosing the study program of an educational institution. After that he or she chooses with AA the proper courses which are related to the chosen program and creates a study plan. Information about the student, the required courses and progress in them is stored in the database and is automatically changed/refreshed after each passed course. After passing several courses, system can attach a student to a group of students with similar, actual study path. If learners are doing well, evidence-based determination and communication of this during advising encourages them to continue like that. If they are attached to a profile which does not progress with the studies as expected, the system can identify this early and provide intrusive academic advising support for both the advisor and the students in question.

The proposed automated mechanism solve an important problem of improving and providing academic advising, because more and more students should receive guidance with their study plans before graduating. This system will help to plan when and how to provide the courses, especially the compulsory ones, as well as to plan a profile intervention adaptively for a larger pool of students, which will reduce the human effort of academic advising.

5 CONCLUSIONS

Academic Advising is an essential part of daily activities in an educational institution and an important component in the learner's study life. Nowadays, we need to be able to create and manage personalized study plans and study paths taking into account learners abilities and regulations of the learning environment. And in order to better help students, Academic Adviser should be able to manage a rich set of information, e.g., on short-range program planning, evaluation of students, and generation of the proper teaching schedule, as well as plan possible interventions adaptively for a group of students instead of following all individual students separately. It is decisive that learner should receive proper advising – poor or no advising is known to have a negative effect on the progress in studies (Al-Ansari et al., 2015).

In this paper, we presented a compact literature review about Academic Advising, mostly focusing on Automated Academic Advising and Intelligent Academic Advising. It was then described how, by using a robust variant of prototype-based clustering method, which is especially suitable for data with missing values, one can create prototypical student group profiles characterizing the overall progress of the studies. This allows academic advisers to provide evidence-based information on the study paths that were actually realized by individual students. Moreover, academic institutions can focus on management and updates on course schedule having an effect on clearly characterized and recognized groups of students. Note that even if the sample groups of students that were profiled here were very small, the used method is scalable to hundreds of thousands of students (Saarela and Kärkkäinen, 2015b).

Then a reference model for automated Academic Advising system was proposed. The proposed architecture and model of the system are intended for a development phase to prototype the whole automated process, where the learners will be profiled regularly, and where the proper study path will be presented, as well as deviating learners detected. The proposed

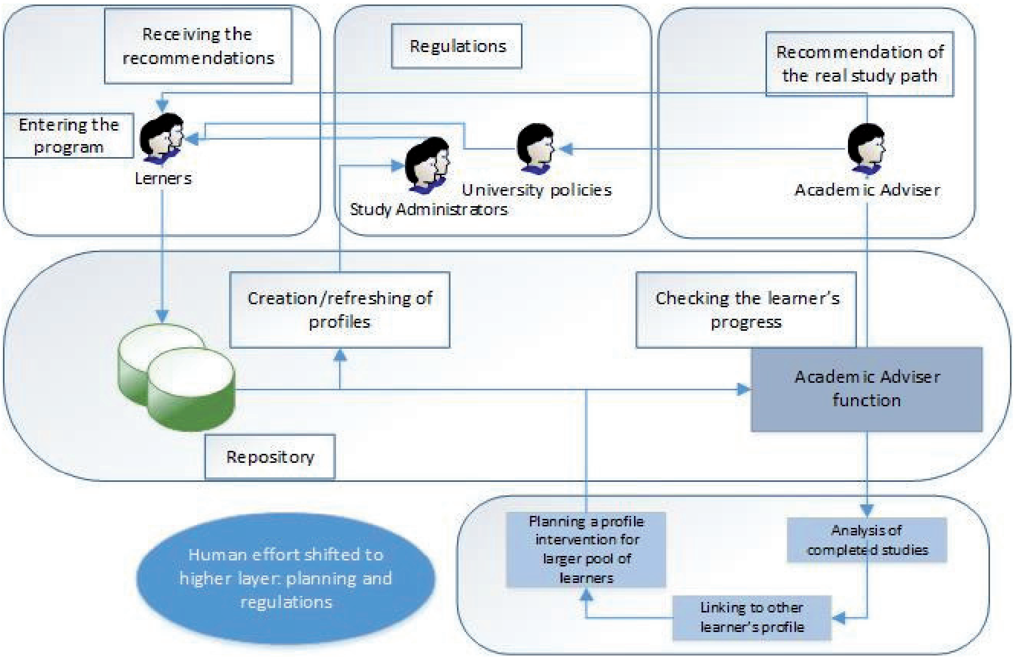

Receiving the recommendations
Regulations
Recommendation of the real study path
Entering the program
Lerners
Study Administrators
University policies
Academic Adviser
Creation/refreshing of profiles
Checking the learner's progress
Academic Adviser function
Repository
Human effort shifted to higher layer: planning and regulations
Planning a profile intervention for larger pool of learners
Analysis of completed studies
Linking to other learner's profile

Figure 7: Automated Academic Advising process.

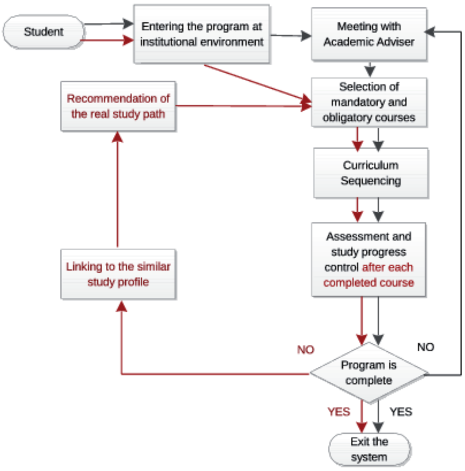

Student
Entering the program at institutional environment
Meeting with Academic Adviser
Recommendation of the real study path
Selection of mandatory and obligatory courses
Curriculum Sequencing
Assessment and study progress control after each completed course
Linking to the similar study profile
NO
NO
Program is complete
YES
YES
Exit the system

Figure 8: Comparison of the manual and automated processes of learner's study life cycle.

model of the AA system will have automated process of study path recommendation. This system will help to plan when and how to provide the courses, especially the compulsory ones, as well as to plan a profile intervention adaptively for a larger pool of students, which will reduce the human effort of academic advising.

By continuing the development of the line of work, we could consider the study paths with higher granularity than per semester. Also, the main functionality of the proposed system – to provide an automated notification for the academic advisers about students and their progress, with the interpretation of needs to modify and re-plan the study path – should be properly evaluated. Moreover, better availability of learner's personal information concerning the study entrance criteria and current life situation, e.g., a part-time job or living far from the institute, could support both interpretation of the generated student profiles and better preparation and management of the intervention patterns of academic advising.

REFERENCES

Al-Ansari, A., El Tantawi, M., AbdelSalam, M., and Al-Harbi, F. (2015). Academic advising and student support: Help-seeking behaviors among Saudi dental undergraduate students. *The Saudi dental journal*, 27(2):57–62.

Al-Sarem, M. (2015). Building a decision tree model for academic advising affairs based on the algorithm C 4-5. *arXiv preprint arXiv:1511.04026*.

Auvinen, T. (2015). *Educational Technologies for Supporting Self-Regulated Learning in Online Learning Environments*. Aalto University.

Auvinen, T., Paavola, J., and Hartikainen, J. (2014). STOPS: a graph-based study planning and curriculum development tool. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*, pages 25–34. ACM.

Äyrämö, S. (2006). *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä.

Biletskiy, Y., Brown, J. A., and Ranganathan, G. (2009). Information extraction from syllabi for academic e-Advising. *Expert Systems with Applications*, 36(3):4508–4516.

Brusilovsky, P. (1998). Adaptive educational systems on the world-wide-web: A review of available technologies. In *Proceedings of Workshop" WWW-Based Tutoring" at 4th International Conference on Intelligent Tutoring Systems (ITS'98), San Antonio, TX*.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., and Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5-6):318–331.

Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, 51(2):787–814.

Chen, C.-M. (2009a). Ontology-based concept map for planning a personalised learning path. *British Journal of Educational Technology*, 40(6):1028–1058.

Chen, C.-M. (2009b). Personalized e-learning system with self-regulated learning assisted mechanisms for promoting learning performance. *Expert Systems with Applications*, 36(5):8816–8829.

Chen, C.-M. et al. (2008). Ontology-based concept map for planning personalized learning path. In *IEEE International Conferences on Cybernetics & Intelligent Systems*.

Chen, C.-M., Lee, H.-M., and Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, 44(3):237–255.

Dougherty, S. B. (2007). Academic advising for high-achieving college students. *Higher Education in Review*, 4:63–82.

Drezner, Z. and Hamacher, H. W. (2001). *Facility location: applications and theory*. Springer Science & Business Media.

Drozd, D. S. (2010). *Student preferences for academic advisors as transformational leaders*. PhD thesis, Texas A&M University.

Feghali, T., Zbib, I., and Hallal, S. (2011). A web-based decision support tool for academic advising. *Educational Technology & Society*, 14(1):82–94.

Fung, A. and Yeung, J. (2000). An object model for a web-based adaptive educational system. In *Proceedings of the IFIP International Conference on Educational Use of Technologies (ICEUT2000)*.

Henderson, L. K. and Goodridge, W. (2015). AdviseMe: An intelligent web-based application for academic advising. *International Journal of Advanced Computer Science and Applications*, 6(8):233–243.

Huang, M.-J., Huang, H.-S., and Chen, M.-Y. (2007). Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems with Applications*, 33(3):551–564.

Huber, P. J. (2011). *Robust statistics*. Springer.

Hwang, G.-J., Kuo, F.-R., Yin, P.-Y., and Chuang, K.-H. (2010). A heuristic algorithm for planning personalized learning paths for context-aware ubiquitous learning. *Computers & Education*, 54(2):404–415.

Idris, N., Yusof, N., Saad, P., et al. (2009). Adaptive course sequencing for personalization of learning path using neural network. *Int. J. Advance. Soft Comput. Appl*, 1(1):49–61.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

Kaihlavirta, A., Isomöttönen, V., and Kärkkäinen, T. (2015). A self-ethnographic investigation of continuing education program in engineering arising from economic structural change. *Studies in Continuing Education*, 37(1):99–114.

Latorre-Navarro, E. M. and Harris, J. G. (2015). An intelligent natural language conversational system for academic advising. *Editorial Preface*, 6(1).

Lee, M.-G. (2001). Profiling students' adaptation styles in web-based learning. *Computers & Education*, 36(2):121–132.

Leung, C. M., Tsang, E. Y., Lam, S., and Pang, D. C. (2010). Intelligent counseling system: A 24 x 7 academic advisor. *Educause Quarterly*, 33(4):n4.

Liu, C.-l., Wu, S., Chang, M., and Heh, J. (2008). Guiding students to do remedial learning in school campus with learning objects' spatial relations. In *16th international conference on computers in education conference*, pages 249–256. Citeseer.

Minaei-Bidgoli, B. (2004). *Data mining for a web-based educational system*. PhD thesis, Michigan State University.

Mostafa, L., Oately, G., Khalifa, N., and Rabie, W. (2014). A case based reasoning system for academic advising in egyptian educational institutions. In *2nd International Conference on Research in Science, Engineering and Technology (ICRSET2014) March*, pages 21–22.

Nguyen, T. B., Nguyen, D. N., Nguyen, H. S., Tran, H., and Hoang, T. A. D. (2008). An integrated approach for an academic advising system in adaptive credit-based learning environment. *Journal of Science, Natural Sciences and Technology*, (24):110–121.

Noaman, A. Y. and Ahmed, F. F. (2015). A new framework for e academic advising. *Procedia Computer Science*, 65:358–367.

Papanikolaou, K. A., Grigoriadou, M., Magoulas, G. D., and Kornilakis, H. (2002). Towards new forms of knowledge communication: the adaptive dimension of a web-based learning environment. *Computers & Education*, 39(4):333–360.

Saarela, M. and Kärkkäinen, T. (2014). Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68.

Saarela, M. and Kärkkäinen, T. (2015a). Analysing Student Performance using Sparse Data of Core Bachelor Courses. *JEDM-Journal of Educational Data Mining*, 7(1):3–32.

Saarela, M. and Kärkkäinen, T. (2015b). Do Country Stereotypes Exist in PISA? A Clustering Approach for Large, Sparse, and Weighted Data. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 156–163.

Saarela, M. and Kärkkäinen, T. (2015). Weighted Clustering of Sparse Educational Data. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015)*, pages 337–342.

Saarela, M. and Kärkkäinen, T. (2017). Knowledge Discovery from the Programme for International Student Assessment. In Peña-Ayala, A., editor, *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*, pages 229–267. Springer International Publishing, Cham.

Sandvig, J. and Burke, R. (2005). Aacorn: A cbr recommender for academic advising. Technical report, Technical Report TR05-015, DePaul University.

Schiaffino, S., Garcia, P., and Amandi, A. (2008). eteacher: Providing personalized assistance to e-learning students. *Computers & Education*, 51(4):1744–1754.

Sprent, P. and Smeeton, N. C. (2016). *Applied nonparametric statistical methods*. CRC Press.

Study Amanuensis (November 2, 2016). private communication.

Tang, T. Y. and McCalla, G. (2005). Smart recommendation for an evolving e-learning system: Architecture and experiment. *International Journal on elearning*, 4(1):105.

Tuttle, K. N. (2000). Academic advising. *New directions for higher education*, 2000(111):15–24.

Wallden, L. J. (2016). *Kansainvälisten koulutusarvioiden vertailu koulutuksellisen tiedonlouhinnan keinoin*. Jyväskylä Studies in Computing. University of Jyväskylä.

Wen, F. L. S. L. D. and McGreal, F. Z. K. R. (2015). e-Advisor: A multi-agent system for academic advising. *International Journal of Advanced Computer Science and Applications*, 6(8).

Werghi, N. and Kamoun, F. K. (2009). A decision-tree-based system for student academic advising and planning in information systems programmes. *International Journal of Business Information Systems*, 5(1):1–18.

Zaki, M. J. and Meira Jr, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

PX

EXPERT-BASED VERSUS CITATION-BASED RANKING OF SCHOLARLY AND SCIENTIFIC PUBLICATION CHANNELS

by

Mirka Saarela, Tommi Kärkkäinen, Tommi Lahtonen, Tuomo Rossi 2016

Journal of Informetrics, 10(3):693–718

Journal of Informetrics 10 (2016) 693–718

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Expert-based versus citation-based ranking of scholarly and scientific publication channels

Mirka Saarela*, Tommi Kärkkäinen, Tommi Lahtonen, Tuomo Rossi

Department of Mathematical Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form 15 February 2016
Accepted 21 March 2016

Keywords:
Journal evaluation
Performance-based funding
Scopus
Journal citation reports
Finnish ranking system

ABSTRACT

The Finnish publication channel quality ranking system was established in 2010. The system is expert-based, where separate panels decide and update the rankings of a set of publications channels allocated to them. The aggregated rankings have a notable role in the allocation of public resources into universities. The purpose of this article is to analyze this national ranking system. The analysis is mainly based on two publicly available databases containing the publication source information and the actual national publication activity information. Using citation-based indicators and other available information with association rule mining, decision trees, and confusion matrices, it is shown that most of the expert-based rankings can be predicted and explained using automatically constructed reference models. Publication channels, for which the Finnish expert-based rank is higher than the estimated one, are mainly characterized by higher publication activity or recent upgrade of the rank. Such findings emphasize the importance of openness of information on a ranking system, with its multifaceted evaluation.

1. Introduction

The quality or impact of a publication channel (i.e., source of publications) can be used for many purposes. Traditionally, the impact of a serial has been used to determine the most important sources of disciplinary knowledge to be acquired for the university libraries – nowadays in digital form. Another, more recent function is to use the research output of universities to evaluate their operational performance through a Performance-based Research Funding System (PRFS). Currently, in many countries, PRFSs have a prominent role in national resource allocation (Abramo & D'Angelo, 2015; Auranen & Nieminen, 2010; Fairclough & Thelwall, 2015). According to Hicks (2012), a PRFS can utilize either an evaluation-based (peer-review) or an indicator-based (bibliometric) model. The prime example of the evaluation-based model was the emergence of the *Research Assessment Exercise* in 1986 and its transformation to *Research Excellence Framework* in England (Wilsdon et al., 2015). For indicator-based models, which are of the main interest here, one has witnessed a transition from the raw numbers of different kinds of publications (e.g., books, articles, and reports) towards their aggregated quality indicators (Haustein & Larivière, 2015). Here an important lesson comes from the *Composite Index* (CI) that was implemented in Australia in 1995, where university funding was based only on the number of publications. However, as shown by Butler (2003), this mostly led to a higher publishing activity in lower quality journals so that the overall impact of the publications dropped. As a

* Corresponding author.
E-mail addresses: mirka.saarela@jyu.fi (M. Saarela), tommi.karkkainen@jyu.fi (T. Kärkkäinen), tommi.j.lahtonen@jyu.fi (T. Lahtonen), tuomo.j.rossi@jyu.fi (T. Rossi).

http://dx.doi.org/10.1016/j.joi.2016.03.004

result, the national PRFS (*Excellence for Research in Australia 2012*) uses both indicators and peer evaluation by an evaluation committee (Vanclay & Bornmann, 2012).

National allocation of research funding using solely an indicator-based model is not common (Hicks, 2012). The PRFS in Flanders (Belgium), as depicted in Verleysen, Ghesquière, and Engels (2014), provides one example, where one of the four pillars of funding for the Flemish universities is based on publications and citations. The Italian research assessment exercise (*Valutazione della Qualita' della Ricerca*) first applied a hybrid peer-review/bibliometrics method during 2004–2010 (Giovanni, Tindaro, & D'Angelo, 2014), and in 2011, introduced a model in which universities were free to choose between peer-reviews and bibliometric indicators as their research evaluation method (Cattaneo, Meoli, & Signori, 2014). The research funding evaluation methodology in Czech (*Metodika hodnocení*) counts all research outputs – among them publications – and then uses aggregated research output points as the basis for the university funding (Good, Vermeulen, Tiefenthaler, & Arnold, 2015). Generally in Europe, as recently summarized by Pruvot, Claeys-Kulik, and Estermann (2015), an output-oriented funding formula as the primary mechanism for research funding is used in England, Finland, Flanders, Ireland, and Poland.

The Nordic system, together with that of Flanders, is distinguished from the other indicator-based PRFS models by the development of open, full coverage national databases in order to record and validate academic publication activity (Verleysen et al., 2014). These databases provide the first basic element of the so-called *Norwegian Model* (NM) that has been described by Ahlgren, Colliander, and Persson (2012), Sivertsen (2010), and Schneider (2009). The main purpose of the NM is to combine (assess) production and quality of publications, without directly using citations. The purpose of the other main components of the model is to create a unified ranking system among various academic disciplines. Finally, the publication points counted using the aggregated ranks determine the university's share in annual government research funding. According to a recent evaluation by Schneider, Aagaard, and Bloch (2015), the NM has proved to serve its purposes in Norway. In particular, in comparison with the above mentioned CI in Australia, the quantity of publications has grown, while the overall quality of publications remained basically the same (Ahlgren et al., 2012; Schneider et al., 2015).

The other two Nordic countries – first Denmark (Schneider, 2009) in 2009, and then Finland (Puuska, 2014, pp. 81–83) in 2010 – have introduced their national PRFSs that follow the NM. Similarly to Norway, the main reason to creating a unified national ranking system in Finland for all relevant publication channels was the difficulty in using available quality indicators to compare the various research and publication cultures of different disciplines (e.g., comparing humanities or social science (SSH) to technology or natural science). The purpose of the Finnish database, *JuFo*[1] is to highlight for the national scientific community the characteristics of all relevant publication channels. Currently, 13% of public university funding in Finland is based on the average weighted sum of quality ranks of all the publications that were produced over a period of three years. The national goal is to target research activity in prestigious international forums, and to enable national evaluation and management of research activities and quality over the years. Hence, *JuFo* serves in Finland both as an available indicator of the quality of publication channels and as a guideline for allocating funding to its national research institutions.

Generally, the quality of a publication channel can be evaluated by an expert in that channel's area of academia (expert-based), or by citation-based indicators of scientific impact (Ahlgren et al., 2012; Ahlgren & Waltman, 2014). The classifications of publication channels in *JuFo* – i.e. the Finnish ranks – are expert-based, like they generally are in the NM as well. Though citation-based indicators can be used as an aid in the NM, the final decisions about the ranks should be made by experts (Sivertsen, 2010). In February 2015, *JuFo* incorporated 29,443 different publication channels, assigning every journal and conference proceeding publication channel to one of 24 expert panels. Each of these 24 panels is composed of experienced and respected Finnish researchers in different scientific fields (all fields can be found in Table A.12). A steering committee allocates publication channels to the panels and provides common ranking rules.

Although the PRFSs of the three Nordic countries following the NM are fairly similar, some crucial differences exist. The Danish and Norwegian PRFSs have the same number of quality ranks: 0 (non-scientific publication channel), 1 (scientific publication channel), and 2 (publication channel with especially great scientific prestige). In both countries, the ranks are updated annually. Publication channels at rank 2 can, at most, account for 20% of the world's publications in a discipline. In Finland, each expert panel must classify all assigned publication channels to one quality category. However, unlike the Norwegian and Danish PRFSs, in Finland, the number of publication channels (not the number of publications) is used to define the quality ranks percentages. Moreover, the Finnish *JuFo* system has one additional rank, (3), which is reserved for the top (at most 5%) of the rank 2 publication channels from each discipline. An additional difference is that in Finland, the ranks of all publication channels in the list are reevaluated only every fourth year. The last reevaluation of all publication channels took place during 2014, and were available in the *JuFo* list in early 2015.

The purpose of this paper is to analyze the expert-based ranks in the *JuFo* list. At the moment, the state covers all costs associated with the publication forum, its management, and the evaluating panels. Furthermore, as argued by the Danish Centre for Studies in Research and Research Policy (2014), one weakness of an indicator like the *JuFo*-rank is the lack of transparency in the nomination process of the steering committee and the panels. As Serenko and Dohan (2011) discovered, an expert's current research interest can strongly influence his or her ranking of publication channels. Therefore, our basic research questions are:

[1] *JuFo* is the abbreviation of "Julkaisusfoorumi", which means "publication forum" in Finnish.

(*i*) Do we need the system in its current form, or can the ranks be automated by rules using available information?
(*ii*) Can possible deficiencies in rankings be linked to certain characteristics of the decision-making process or the decision makers?

The similarity of the Finnish system to that of Norway and Denmark implies that the present study compares expert judgment on the individual publication channel level across three countries, which involves, as far as we know, a novel research setting. As a result, the main question of the present study – following the research track of the previous studies by Ahlgren et al. (2012), Ahlgren and Waltman (2014), and the Danish Centre for Studies in Research and Research Policy (2014) – is to address the two basic approaches for evaluating publication quality: expert-based versus citation-based publication channel ranking.

We propose to answer the research questions by linking two publicly available national datasets to external reference measures retrieved from Thomson Reuters' Journal Citation Reports® and Scopus®. We argue that repeatable patterns and rules, based on available relevant information, can be used to modify the entire ranking system, using central decision making to improve ranking efficiency and transparency. Thus, automatic and repeatable rules would help to open up the nomination process, assist panel members in their decision-making, and possibly save work and costs related to the system. Notably, studies that discuss automatization of expert judgment in research evaluation on the basis of advanced methodology and large datasets presently have a broad interest in research policy (Wilsdon et al., 2015).

In the existing quantitative evaluations of expert-based rankings (e.g., Ahlgren et al., 2012; Ahlgren & Waltman, 2014; Vanclay, 2011), typically a small set of citation-based indicators are linked to the expert-based rankings. This means that only those publication channels that have a reference citation-based indicator can be assessed. Here, our goal is to enlarge (even maximize) the coverage of each expert-rank evaluation by incorporating into it more explanatory variables (metadata) and involved data analysis techniques than existing studies do. For instance, the binary information concerning whether or not certain citation-based indicators are available has clear relationship to the ranking, as will be evident later on. Therefore, our contributions are two-fold: First, we address the broad international relevance concerning the question of expert-based versus citation-based publication channel rankings. Second, we use a novel methodological approach combined with available data from large datasets to analyze the expert-based decisions.

The structure of this paper is as follows: First, we describe the *JuFo* data and its available attributes (Section 2). Second, we present our overall analysis method (Section 3), which is based on triangulated (Bryman, 2004) machine learning techniques. Third, we present how well the rules we identified can predict the Finnish expert-based ranking (Section 4). Moreover, we characterize the publication channels that are misclassified when the aforementioned rules are used. Finally, overall patterns and findings are presented in Section 5.

2. Data

The data for this study comes from three sources:

1. *JuFoDB*: Database of the Finnish publication forum, *JuFo*[2], which contains all nationally evaluated publication channels. Data was retrieved from this database in February 2015.
2. *JuuliDB*: The publicly accessible database of *Juuli*[3] that contains all publications of Finnish researchers. Each publication channel in *JuFoDB* has a unique *Juuli* ID, through which all Finnish publications in that particular channel can be found. Data was retrieved from this database in September 2015.
3. The *Journal Citation Reports*® (JCR): Published by Thomson Reuters, there is no direct link available from *JuFoDB* to the JCR. However, 8178 of all the 8539 observations from the Thomson Reuters database were linked to publication channels in *JuFoDB* by using the ISSN available in both databases. Data from this database was retrieved in September 2014.

Altogether, 29,443 different publication channels with 33 attributes were retrieved from *JuFoDB*. The example in Table 1 shows available attributes for the *Journal of Informetrics*. As can be seen from the table, the *Journal of Informetrics* has been evaluated as one of the most prestigious journals in its field (rank 3). The Finnish expert-based rank (i.e., the *JuFo*-level) of each publication channel as well as the Norwegian and Danish expert-based rankings can be obtained directly through the *JuFoDB*. Moreover, as can be seen in Table 1, the three indicators from the bibliographic database Scopus®, that is the SJR, the SNIP and the IPP, are featured. Furthermore, by using the ISSN linkage to Thomson Reuters' JCR, we can, for the common publication channels, access the six original JCR variables (*Total Cites*, *Articles*, *Impact Factor*, *Cited Halflife*, *Immediacy Index*, and *5-Year Impact Factor*), as well as the two Eigenfactor metrics (*Eigenfactor Score* and *Article Influence Score*).

In addition to some more general data, such as the unique identifier (*ID*), *ISSN*, and *publisher*, the *JuFoDB* also provides the information on the panel (see Table A.12) responsible for evaluating a publication channel. Moreover, through the link to *JuuliDB* (the last attribute in Table 1), one can directly access the information of all researchers in Finland who have published in the particular channel. Additional information such as *abbreviation*, *ISBN*, *end year*, *continued under the name* and *continued JuFo-rank* are available for some publication channels, but were not included in Table 1 because they were not available for

[2] Available at: http://www.tsv.fi/julkaisufoorumi/haku.php.
[3] Available at: http://www.juuli.fi/?&lng=en.

Table 1
Information available in *JuFoDB* with the example of the Journal of Informetrics.

Attribute	Value
Level	3
JuFo ID	60692
Title	Journal of Informetrics
Parallel title or subtitle	
Title details	
Website	http://www.journals.elsevier.com/journal-of-informetrics/
Type	Serial
ISSN (print)	1751-1577
ISSN (online)	1875-5879
Starting year	2007
Country of publication	NETHERLANDS
Publisher	Elsevier BV
Language	English
Norway level	1
Denmark level	2
ERIH field	
SJR	2.541
SNIP	2.018
IPP	3.51
DOAJ	No
Sherpa/Romeo	Green
Evaluating panel	17
Field	;1 Natural Science;5 Social Science;
MinEdu field	111 Mathematics 112 Statistics 113 Computer and information sciences 512 Business and management 518 518 Media and communications
Web of Science fields	INFORMATION SCIENCE &LIBRARY SCIENCE (SSCI)
Scopus fields	Modelling and Simulation Management Science
Evaluation history	Level 2015: 3 Level 2014: 2 Level 2013: 2 Level 2012: 2
Juuli	60692

Table 2
Comparison of the discipline-wise rankings in *JuFo*.

Discipline	ID	Rank 0	Rank 1	Rank 2	Rank 3	Total occurrences
Natural science	1	1243 (14%)	6628 (75%)	733 (8%)	248 (3%)	8852 (100%)
Technology	2	1294 (26%)	3300 (65%)	348 (7%)	100 (2%)	5042 (100%)
Medical and health	3	250 (5%)	4615 (85%)	430 (8%)	113 (2%)	5408 (100%)
Agriculture and forestry	4	106 (10%)	904 (83%)	61 (6%)	24 (2%)	1095 (100%)
Social science	5	1521 (18%)	5777 (69%)	865 (10%)	267 (3%)	8430 (100%)
Humanities	6	652 (9%)	5196 (75%)	837 (12%)	219 (3%)	6904 (100%)
Other	9	22 (27%)	52 (64%)	1 (1%)	6 (7%)	81 (100%)
All disciplines	–	5088 (14%)	26,472 (74%)	3275 (9%)	977 (3%)	35,812 (100%)

the *Journal of Informetrics*. None of the observations in the database is complete, meaning that all of the publication channels have missing values for at least some of the 33 total attributes. Hence, for utilizing all of available data in the analysis, one faces a significant sparsity problem (see, e.g., Saarela & Kärkkä inen, 2015, and articles therein).

Each publication channel in *JuFoDB* is assigned to at least one discipline. Rankings are presented according to discipline in Table 2, in which most of the publication channels have been evaluated as basic, or rank 1. We can see from the table that the percentages do not differ much between the disciplines. However, Natural Science has more than ten times more rank 2 and 3 publication channels than Agriculture and Forestry. Even if this should reflect the size of the overall publication channel population, it probably better reflects the size of the national researcher population in a discipline. From both a discipline and panel perspective, the more publication channels can be brought under evaluation, the more high ranks can be given in absolute terms.

Table A.12 provides information about the distribution of the different ranks according to the panels. Although a discipline may have multiple possible linkings, each publication channel is attached to only one panel. As can be seen in the table, some differences exist when it comes to the percentage of the highest classified publication channels across the panels. However, all panels adhere to the rule (see Section 1) that 20% of the publication channels at most are allowed to be classified as rank 2, and 5% at most as rank 3. There is no panel information available in *JuFoDB* for 6562 observations (see the first column

Table 3
Overview of used variables, their availabilities in percentages, and preprocessings.

Variable	Availability	Preprocessing
rank	100%	The categorical (0–3) Finnish expert-based rank (*JuFo*-rank).
panel	100%	The categorical indicator which panel (1–24) was responsible for evaluating the channel.
type	100%	3 for journals, 2 for conferences, 1 for book publishers.
inJCR	100%	1 if the publication channel can be found in Thomson Reuters' JCR, 0 otherwise.
nrOfPub	100%	The total number of publication in this channel as retrieved from *JuuliDB*.
rankChange	100%	0 if there was no change compared to the rank in the previous year, 1 if the current rank is lower, and 2 if the current rank is higher than in the previous year.
language	94.76%	3 for English, 2 for Finnish or Swedish, 1 for other languages. NaN if not available.
age	91.12%	Current year (2015) minus the start year of the channel. NaN if start year cannot be found.
NORrank	76.56%	The categorical (0–2) Norwegian expert-based rank. NaN if not available.
DNKrank	71.0%	The categorical (0–2) Danish expert-based rank. NaN if not available.
SJR	67.81%	The continuous SJR value. NaN if not available.
SNIP	65.06%	The continuous SNIP value. NaN if not available.
IPP	64.69%	The continuous IPP value. NaN if not available.
sherpaCode[1]	60.10%	5 for green, 4 for blue, 3 for gray, 2 for yellow, 1 for white. NaN if not available.

[1] Definitions and terms of Sherpa are provided at http://www.sherpa.ac.uk/romeoinfo.html.

in Table A.12). The publication channels that have not been assigned to any panel have a special profile: they are all book publishers and have mostly been evaluated as rank 0 (see Table A.12). What is not clear, based on the general description of the ranking system as described in Section 1, is who could update the ranking of the non-panel-allocated publication channels?

2.1. Observations and variables used for the study

For further analysis, we selected all 22,881 observations from *JuFoDB* that were assigned to a panel. Moreover, we utilized all available variables that might affect the expert rank of a publication channel. Table 3 provides an overview of all the used variables, their preprocessings, and their availability with respect to the 22,881 observations under study. It is important to note that the distribution of ranks (0–3) of our observations is very imbalanced: 2.01% are rank 0, 2.87% are rank 3, 9.85% are rank 2, and 85.27% are rank 1. Hence, for example a trivial classifier returning always 'rank 1' would be more than 85% correct. As will be seen below, by using the proposed methods and techniques, we only obtain slightly better overall classification accuracies than this. However, compared to the trivial classifier, the advantage of these methods and techniques is their explicit construction allowing one to identify and discuss salient variables of the models.

3. Method

Our analysis was based on a combination of different machine learning techniques (e.g., Alpaydin, 2010) with a unified analysis pattern: We first generated an automatic indication of ranks and, then, studied the deviations from this to analyze their characteristics. All computations were performed using Matlab 2015b. The applied techniques and deviations used are as follows:

- Association rules to determine patterns in the data based on the availability of variables (deviations are defined as publication channels for which the rules do not apply)
- Decision tree with stratified cross-validation to construct a classification model for the ranks, using the through association rules detected patterns in data (deviations are defined as misclassified publication channels)
- Reference indicator detection using triangulated PCA for Thomson Reuters' JCR (confusion matrices are used to define deviations from the baseline)

3.1. Decision tree

We aim to predict the Finnish expert-based ranking by automatic rules. Decision tree is a supervised machine learning technique that can predict the categorical output (rank) from given categorical and continuous predictor variables. It is very suitable in our case because we are interested in a prediction model that provides explicit rules with respect to the predictor variables used. A decision tree presents the rules in a tree-like structure, whose nodes provide readable and easily accessible rules on the so-called splitting variable for human interpretation. We use the CART (Breiman, Friedman, Stone, & Olshen, 1984) decision tree induction algorithm (default in Matlab), in which the splitting is based on Gini's diversity index.

However, one problem with using a decision tree (explicitly visible in Table 3) is the high percentage of missing values in data. Observations that have a missing value for a splitting variable are automatically assigned to the most frequent class. This is especially unsuitable in our case, since we have (as already discussed in Section 2.1) very imbalanced class sizes. If we use a decision tree for the whole data, we receive an almost perfect classifier. However, this is not because the classifier

Table 4
Confusion matrix to identify highly deviating publication channels.

Ref. rank	*JuFo*-rank			
	Rank 0	Rank 1	Rank 2	Rank 3
Rank 0	+	+	–	–
Rank 1	+	+	+	–
Rank 2	–	+	+	+
Rank 3	–	–	+	+

itself has built a valuable model that captures the data very well. Instead, every time a variable with missing values is used as the splitting attribute, the classifier can assign all observations with missing values for this variable to the most common class (i.e., rank 1, which by default is in more than 85% of the cases correct).

We solve this sparsity problem by using association rules indicating for each variable whether the value is missing or not (see Section 3.2). Furthermore, we solve the problem of the imbalanced class sizes in the decision tree by assigning the inverse of its class frequency to each observation as a weight. This technique is called oversampling (He & Garcia, 2009).

3.2. Association rules

The goal of association rule mining (Agrawal, Imieliński, & Swami, 1993) is to automatically find patterns that describe strongly associated attributes in data. The discovered patterns are usually represented in the form of implication rules or attribute subsets. If I is the set of all items and S_1 a subset of the set of items ($S_1 \subseteq I$), a transaction $t_i \in T$, where T denotes the set of all transactions, is said to contain itemset S_1 if S_1 is a subset of t_i. The support count, $\sigma(S_1)$, for an itemset S_1 is defined as $\sigma(S_1) = |\{t_i \mid S_1 \subseteq t_i, t_i \in T\}|$, where $|\cdot|$ stands for the cardinality, i.e., the number of elements in a set.

An association rule is then an implication expression of the form $S_1 \rightarrow S_2$, where $S_1, S_2 \subseteq I$ and $S_1 \cap S_2 = \emptyset$. The support, $s(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{|T|}$, determines how often a rule is applicable to a given data set. Furthermore, the confidence, $c(S_1 \rightarrow S_2) = \frac{\sigma(S_1 \cup S_2)}{\sigma(S_1)}$, determines how frequently items in S_2 appear in the transactions that contain S_1.

Association rule mining is applied to the whole data set, i.e., to all 22,881 observations under study. Our itemsets consist of binary representation (encoding) of all the variables presented in Table 3, except the number of publications and the rank change, as those should not have an effect on the expert-based rank. Hereby, we use for all categorical variables in each case one variable for each category, and, if there can be missing values, one additional variable indicating whether the value is missing. For example, with this strategy we have for language the binary indications *isEnglish*, *isFinnishOrSwedish*, *otherLanguage*, and *languageNaN*, while for rank (which is available for all observations), we only have the binary indications *rank 0*, *rank 1*, *rank 2*, and *rank 3*. Furthermore, for our three continuous variables (SJR, SNIP and IPP) and one discrete variable (age), we use two variables (e.g., *SJRavail*, and *SJRnan*) in each case to indicate whether these variables are available or not. Altogether, that gives us 59 binary variables for each observation.

We are interested in association rules with high confidence, as confidence represents the reliability and accuracy of a rule. On the other hand, support can be relatively small, since we are interested in all rules that contain rank information. For example, a transaction that contains the item *rank 0* can by construction be supported by at most 2.01% (see Section 2.1) of all transactions in the itemset.

3.3. Confusion matrix using a reference metric

The idea for our third analysis technique is to compare the existing *JuFo*-rank of each observation in the database to an overall reference indicator, using a simple confusion matrix (Alpaydin, 2010). Thus, the (continuous) reference indicator is categorized to have the same number of ranks as present in *JuFo*. Hereby, we accept small deviations from perfect matches. We entitle the *JuFo*-rank of a publication channel to be in accordance with the reference indicator (denoted as + in Table 4) if the reference indicator rank is either equal to or at most one rank higher or lower than the *JuFo*-rank. Furthermore, we characterize the *JuFo*-rank as highly deviating (denoted as −) if the *JuFo*-rank is at least two ranks higher (or lower, respectively) than the reference indicator. We study further those observations that deviate greatly from the reference indicator, asking, “Which publication channels have been evaluated very differently by the Finnish panels compared to a constructed reference indicator, and can they be summarized by a general profile?”

Defining one overall reference indicator is challenging. Traditionally, the impact factor (IF) published by Thomson Reuters in the JCR (see Section 2) has been the most well-established ranking for the evaluation of publication channels. However, as discussed by numerous scholars before (Archambault & Larivière, 2009; Falagas, Kouranos, Arencibia-Jorge, & Karageorgopoulos, 2008; Moed, 2010; Seiler & Wohlrabe, 2014; Vanclay, 2012), the IF has several limitations, such as the lack of quality assessment of the citations and the influence of journal self-citations. Due to its simple formula (the IF is computed by dividing the number of citations in the JCR yearly by the total number of articles published in the two previous years), a journal can easily boost its IF by accepting only articles that cite a certain percentage of recent articles from the same journal. Furthermore, citation practices differ between disciplines (Moed, 2010) and, as a consequence, the likeliness of being cited depends also on the research field. Reference lists in mathematical articles, for example, tend to be much shorter than those in biology. More precisely, as shown in (Moed, 2005, Chapter 5), the top journals in large disciplines typically have higher citation impact than the top journals of smaller disciplines. Therefore, especially because we intend to evaluate the *JuFo*-rank across different scientific fields using the same indicator, the IF alone cannot serve as our external reference quality indicator.

With the exception of considering another yearly time interval (and for IPP, another database), the *5-Year Impact Factor* (five years), *Immediacy Index* (current year), and IPP (three years) are fairly similar in construction to the traditional IF (two years). Although we cannot use these metrics because of the aforementioned reasons, we note that these still seem to be the most established external citation-based metrics. This is evidenced by the fact that usually, when a new indicator is introduced, it is compared against the IF of some yearly time interval. For example, Moed (2010) compared the SNIP indicator against the three-year IF (IPP[4]), Guerrero-Bote and Moya-Anegón (2012) compared the SJR2 against the three-year IF (IPP), and González-Pereira, Guerrero-Bote, and Moya-Anegón (2010) compared the SJR against the three-year IF (IPP). Furthermore, after the two Eigenfactor® metrics had been introduced by Bergstrom, West, and Wiseman (2008), Franceschet (2010) compared both to the two-year IF, and the five-year IF.

To sum up, despite the fact that all recent studies seem to agree that the IF of some yearly time interval is outdated and not adequate for a fair comparison across disciplines, no other external citation-based metric seems to have reached the same status yet. Bollen, Van de Sompel, Hagberg, and Chute (2009) point out that there is not even a “workable definition of the notion of ‘scientific impact’ itself.” The general conclusion is that no single indicator alone “captures all the criteria that are needed for a rigorous and comprehensive measure of scientific output” (Kreiman & Maunsell, 2011). Instead, the quality of scientific publication channels is a “multi-dimensional concept that cannot be expressed in any single measure” (Moed, 2010).

3.3.1. *Finding an overall reference quality indicator*

Based on our conclusion that the quality of scientific publication channels must be described by multiple features (and possibly several metrics), we used the different quality indicators published by Thomson Reuters as our starting point. Six different JCR metrics were available, plus the two Eigenfactor® metrics (see Section 2). However, two of these eight measures, namely *Cited Halflife* and *Articles* were not directly connected to the quality of a publication channel. Hence, we left out the *Cited Halflife* as “a higher or lower cited half-life does not imply any particular value for a journal”[5]. Furthermore, we decided to omit the *Articles* indicator since it does not necessarily increase the quality of a journal if more articles are published, and does not correlate as strongly with the other variables.

All remaining six indicators in the Thomson Reuters database are positively correlated (see Fig. 1). Hereby, we observe immediately two groups that have especially strong metric correlations with each other. The first group (composed of *Total Cites* and *Eigenfactor Score*) represent the not-normalized metrics (see also Table A.13). The second group (*Impact Factor*, *5-Year Impact Factor*, *Immediacy Index*, *Article Influence Score*) is composed of metrics that normalize the influence of a journal with regard to its publishing volume, i.e., they measure the average influence of an article in the journal. A high correlation between different variables of the JCR data has been observed by other scholars, as well. In Chang, McAleer, and Oxley (2013), a table can be found that provides an overview of which correlation between which variables have been observed by which researchers. Since these six indicators all measure the same concept, i.e., the quality of a publication channel, and our intention was to describe this concept in a compact way (ideally at once), we scaled all six variables to the interval [0, 1] using min-max scaling to prepare them for dimension reduction. Then, we applied *Principal Component Analysis* (PCA). Because we had missing values (see Table A.14, which summarizes the availability for each indicator in the JCR), we had to use PCA for sparse data. To strengthen the result using methodological triangulation (Bryman, 2004), we used three different approaches.

The classical PCA for sparse data, the robust PCA for sparse data (see both Kärkkäinen & Saarela, 2015) and the ALS algorithm (Kuroda, Mori, Masaya, & Sakakihara, 2013) all suggest that the first two principal components explain more than 90% of the variance in the data and that the first three components account for nearly 90% of the geometric variability, respectively. Moreover, the angles between the three principal derived subspaces are very small, which means that the results of the three different PCA variants coincide in practice (cf. Kärkkäinen & Saarela, 2015). Hence, we use the projection of data into the most significant principal (major) component to summarize the six different quality measures in the JCR data.

3.3.2. *Relation of the combined JCR data to the three Scopus indicators*

Next, we study the relation of the SJR, SNIP and IPP indicators to the major component just defined. The three indicators from Scopus offer two main advantages over the quality indicators offered by Thomson Reuters: First of all, the SJR, SNIP and IPP indicators are open-access resources, while access to the JCR data requires a paid subscription. Furthermore, possibly as a consequence of the accessibility, the three indicators can also be obtained directly from the public *JuFoDB*, while the connection to the Thomson Reuters data requires substantially more action (see Section 2). Second, considerably more

[4] Back then, the IPP was known as “Raw Impact per Paper.”

[5] See http://admin-apps.webofknowledge.com/JCR/help/h_ctdhl.htm.

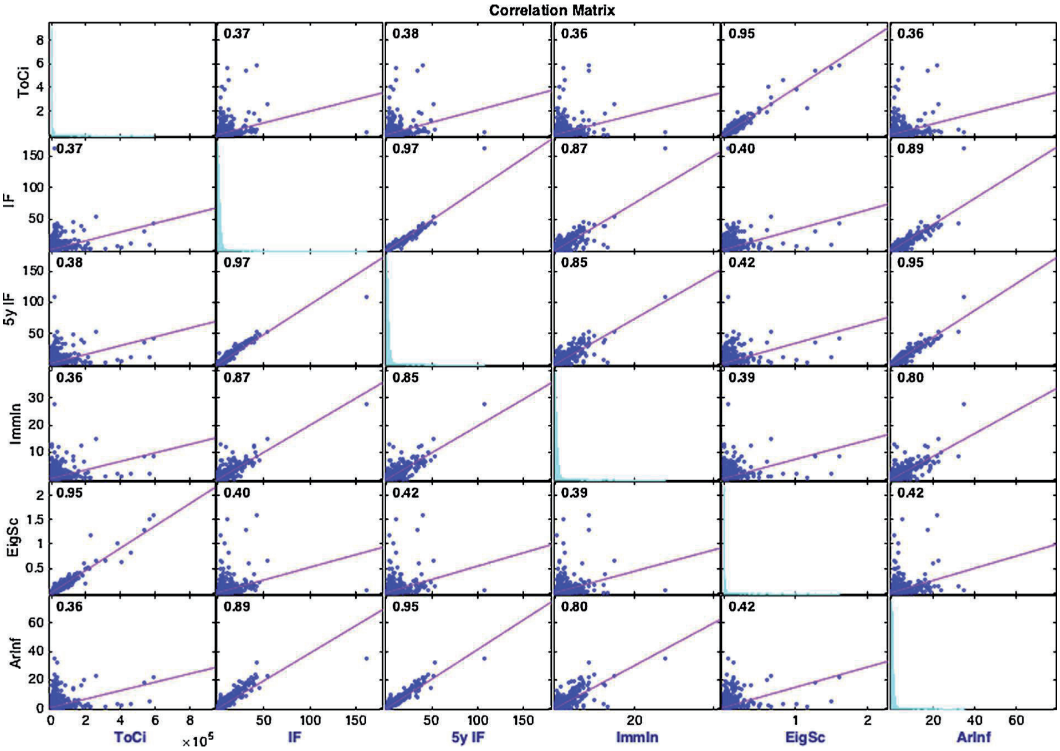


Fig. 1. Spearman correlations of Thomson Reuter measures of quality.

publication channels published in a wider variety of countries and languages are listed in the Scopus database (with SJR, SNIP and IPP values available) than can be found in the Thomson Reuters database (Bornmann et al., 2009; Falagas et al., 2008; Guerrero-Bote & Moya-Anegón, 2012). In our case, only within *JuFo*, 8178 publication channels can be linked to Thomson Reuters JCR, but 17,355 readily have an SJR indicator available.

The correlation between the SJR, SNIP and IPP and each available quality indicator in Thomson Reuters' JCR for those 8178 publication channels that are stored in both databases can be found in Table A.14. As the table shows, the highest correlation ($r = 0.97411$) is observed between the IPP and Thomson Reuters' IF. The second highest correlation ($r = 0.941$) is observed between the SJR and Thomson Reuters' *Article Influence Score* (AI). The very strong correlation of these two metrics is not surprising, given that the SJR is very similar in construction to AI. Both indicators take into account not only the quantity but also (by giving each citation a weight) the quality of the citations using Google's PageRank algorithm. Besides being computed over different databases (Scopus versus JCR), the SJR differs in three additional respects from the AI. First, the SJR is computed over a three-year-window while the AI is computed over a five-year-window. Second, journal self-citations are only limited so that they count for not more than one third of the total citations, while they are totally excluded in the AI. Third, the AI is normalized only by the number of identified references in the citing journals (i.e., those in the JCR data), while the SJR is normalized by the number of all references in the citing journals.

The correlation of SJR with the representation of the Thomson Reuters data spanned by the first principal component using robust PCA is very strong ($r = 0.913$). Likewise, the correlation of the first principal component to SNIP is $r = 0.7398$, and the correlation to IPP is $r = 0.9191$. We conclude that the three Scopus indicators are, with reference to the Thomson Reuters data and with respect to their availability, the most appropriate choices as the citation-based indicators of publication channel quality.

4. Results

4.1. Association rules

We applied association rule mining for the variables and observations, as explained in Section 3.2. Our main interest was in those rules that contained Finnish expert-based rank information. Since we wanted to have rules for all ranks and only

Table 5
Association rules for the whole *JuFo* data.

Rule	Support	Confidence	Rule	Support	Confidence
rank 3 → isJournal	2.87%	100%	SherpaNaN → rank 1	36.77%	92.17%
rank 3 → SJRavail	2.85%	99.24%	SNIPnan → rank 1	32.1%	91.86%
rank 3 → SNIPavail	2.85%	99.24%	IPPnan → rank 1	32.39%	91.73%
rank 0 → notInJCR	1.99%	99.13%	SJRnan → rank 1	29.45%	91.5%
rank 3 → IPPavail	2.84%	98.93%	NOR 1 → rank 1	59.75%	91.43%
rank 0 → IPPnan	1.96%	97.39%	rank 3 → DNK 2	2.62%	91.31%
rank 2 → isJournal	9.57%	97.07%	rank 2 → SJRavail	9%	91.31%
rank 1 → isJournal	80.31%	94.18%	DNKnan → rank 1	26.71%	90.96%
rank 3 → isEnglish	2.67%	93.29%	rank 2 → SNIPavail	8.96%	90.91%
DNK 1 → rank 1	54.69%	93.21%	NORnan → rank 1	21.59%	90.68%
otherLanguage → rank 1	19.77%	92.29%	rank 2 → IPPavail	8.92%	90.55%

a bit more than 2% of all the publication channels under study were *rank 0*, we set the minimal support to 1.95. When the confidence was set to 90% and we explicitly searched for the rules that included the rank information, we obtained the set of rules as presented in Table 5.

As can be seen from Table 5, only one rule is supported 100%. All publication channels evaluated as 3 in *JuFo* were journals. However, also 97% of the publication channels evaluated as 2 in *JuFo*, and 94% of the publication channels evaluated as 1 in *JuFo* were journals. Therefore, the type of publication channel did not seem to be a very useful indicator of the Finnish expert-based rank.

The most interesting subset of the obtained rules was the one that included the availability of reference indicators. We see from Table 5 that if a publication channel has been highly evaluated (i.e., as rank 3 or 2), then the three indicators from Scopus are available with a very high percentage. If the Finnish expert rank is 3, SJR and SNIP are available for more than 99% of all the publication channels, and IPP is available for almost 99% of all the publication channels. If the Finnish expert rank is 2, SJR, SNIP and IPP are available for more than 90% of all publication channels. Vice versa, for more than 97% of those publication channels that have been evaluated as 0, the IPP value is missing. Moreover, we see from the table that of those publication channels missing SNIP, IPP, and SJR, more than 91% have been ranked as 1 in Finland. Hence, it can be concluded that the availability of the three Scopus metrics already provides a very good prediction of the Finnish expert-based rank.

Table 6
Characteristics of misclassified publication channels with association rules.

	SJR, SNIP or IPP missing		SJR, SNIP and IPP missing	
	Rank 3	Rank 2	Rank 3	Rank 2
total sum	7 (0.03%)	214 (0.94%)	5 (0.02%)	196 (0.86%)
in JCR	2 (28.57%)	11 (5.14%)	2 (40%)	9 (4.59%)
mean number of publications	13.29	13.44	14.2	14.32
rank now higher	2 (28.57%)	64 (29.91%)	1 (20%)	55 (28.06%)
no rankChange	5 (71.43%)	138 (64.49%)	4 (80%)	130 (66.33%)
rank now lower	0 (0%)	12 (5.61%)	0 (0%)	11 (5.61%)
Language NaN	0 (0%)	29 (13.55%)	0 (0%)	29 (14.8%)
English	4 (57.14%)	115 (53.74%)	3 (60%)	99 (50.51%)
Finnish or Swedish	0 (0%)	16 (7.48%)	0 (0%)	16 (8.16%)
other Language	3 (42.86%)	54 (25.23%)	2 (40%)	52 (26.53%)
Journal	7 (100%)	148 (69.16%)	5 (100%)	137 (69.9%)
Conference	0 (0%)	66 (30.84%)	0 (0%)	59 (30.1%)
Age NaN	0 (0%)	37 (17.29%)	0 (0%)	37 (18.88%)
Age (mean of avail)	49.57	41.5	39.4	41.41
NOR rank NaN	2 (28.57%)	84 (39.25%)	2 (40%)	82 (41.84%)
NOR rank equals FI rank	5 (71.43%)	52 (24.3%)	3 (60%)	46 (23.47%)
NOR rank higher FI rank	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NOR rank lower FI rank	0 (0%)	78 (36.45%)	0 (0%)	68 (34.69%)
DNK rank NaN	2 (28.57%)	103 (48.13%)	2 (40%)	99 (50.51%)
DNK rank equals FI rank	5 (71.43%)	60 (28.04%)	3 (60%)	57 (29.08%)
DNK rank higher FI rank	0 (0%)	0 (0%)	0 (0%)	0 (0%)
DNK rank lower FI rank	0 (0%)	51 (23.83%)	0 (0%)	40 (20.41%)

Table 7
Characteristics of publication channels that are *JuFo*-rank 2 but have no other available reference indicator.

total sum	73
in JCR	0(0%)
mean number of publications	15.22
rank now higher	22(30.14%)
no rankChange	50(68.49%)
rank now lower	1(1.37%)
Language NaN	29(39.73%)
English	30(41.10%)
Finnish or Swedish	5(6.85%)
other Language	9(12.33%)
Journal	23(31.51%)
Conference	50(68.49%)
NOR rank NaN	73(100%)
DNK rank NaN	73(100%)

4.1.1. Deviations

In Table 6, we summarized the characteristics of those publication channels in which (i) SJR, SNIP or IPP were not available but the Finnish expert-based rank was still 3 or 2 (columns 2 and 3), and (ii) SJR, SNIP and IPP were not available but the Finnish expert-based rank was still 3 or 2 (columns 4 and 5).

As can be seen from the fourth column in Table 6, five publication channels are classified as *JuFo*-rank 3, although SJR, SNIP and IPP are not available. However, three of these have been classified as the highest rank in Norway and Denmark, and for the other two publication channels (namely *British Medical Journal* and *Light: Science & Applications*), the SJR, SNIP and IPP indicators seem to have been incorrectly not included in *JuFoDB*. They can be found by manually using the *Browse Sources* search function in Scopus,[6] and in both cases, the values are so high that rank 3 seems appropriate.

As can be seen from the fifth column in Table 6, 196 publication channels have been classified as *JuFo*-rank 2 but no SJR, SNIP or IPP value is available. However, again for most of these channels, a Norwegian or Danish rank is available and in each case, at least one is classified as rank 1 or higher. Only 73 publication channels remain that are *JuFo*-rank 2 but have no SJR, IPP or SNIP, nor is a Norwegian or Danish expert-based rank available. All of these publication channels are explicitly listed in Table A.15, and a summary is provided in Table 7.

From Table 7, we see already that in the subgroup of those 73 publication channels that have a Finnish expert-based ranking of 2 but no other reference indicator available, a high percentage of conference series exists (68.49%). It is interesting that the number of publications for these channels is very high, with an average of more than 15 per channel. Moreover, for more than one third of the publication channels in this subgroup, the *JuFo*-rank was upgraded during the last evaluation round.

Then, just as above for the *JuFo*-rank 3 publication channels, we tried to manually find SJR, SNIP and IPP values using the search interface in Scopus. Indeed, for 15 publication channels, SJR, SNIP and IPP values were detected, and for 14 others, the coverage discontinued in Scopus (both are reported in Table A.15). If we subtract these 15 + 14 from the 73 publication channels that are *JuFo*-rank 2 but have no other reference indicator available, 44 publication channels remain.

These 44 publication channels show a clear profile: most of them (precisely 26) are conference proceedings evaluated by the *Computer and Information Sciences* panel. The remaining publication channels that do not belong to the *Computer and Information Sciences* panel are journals, mostly published in another language than English. They belong to five different panels: 17,18, 19, 22 and 23. To this end, we also scanned manually through these 44 publication channels and their links to *JuuliDB*. Then, two slightly alarming cases were noticed, in which the highly deviating *JuFo*-rank 2 of a publication channel could be linked to an active publication profile of a panel member responsible for deciding the rank.

However, these results must be interpreted cautiously. As discussed in Section 1, one of the main reasons for establishing the expert-based *JuFo*-ranks instead of using citation-based indicators to measure the quality of a publication channel was to include the SSH and engineering sciences on the same terms as the other major areas. For example, it is known that publication channels belonging to the SSH are less covered in the international citation databases (although there has been a positive trend towards broader coverage, especially in Scopus) than other disciplines (Sivertsen, 2014). Table 8 summarizes for all panels the number and percentages of journals (*JuFo*-rank 0 publication channels as well as book publishers and conference series are excluded) and articles, i.e. publications in *JuuliDB*, within these journals that are covered in Scopus. The table clearly shows the disciplinary differences. While the journals that belong to the panels of the so so-called hard sciences are covered very well in Scopus, the journals that belong to SSH panels are covered relatively less in Scopus. This trend is even more visible when looking at the number and percentages of articles. On average, almost 78% of the total

[6] See http://www.scopus.com/source/browse.url?zone=TopNavBar&origin=searchbasic.

Table 8
Disciplinary differences of journal coverage in Scopus and Finnish publication activity within those (*JuFo*-ranks 0 as well as book publishers and conference series are excluded).

Panel name	Journals			Publications in journals		
	All	In Scopus (% of all)	Not in Scopus (% in FI/SWE)	All	In Scopus (% of all)	Not in Scopus (% in FI/SWE)
Mathematics and statistics	804	644 (80.1%)	160 (0%)	1438	1345 (93.53%)	93 (0%)
Computer and information scis.	652	518 (79.45%)	134 (0.75%)	1874	1705 (90.98%)	169 (1.18%)
Physical scis., space scis. & Astronomy	486	424 (87.24%)	62 (0%)	5136	5023 (97.8%)	113 (0%)
Chemical scis.	444	404 (90.99%)	40 (0%)	2459	2427 (98.7%)	32 (0%)
Geosciences & environmental scis.	624	545 (87.34%)	79 (1.27%)	2028	1967 (96.99%)	61 (0%)
Biosciences I	621	573 (92.27%)	48 (4.17%)	2839	2744 (96.65%)	95 (49.47%)
Biosciences II	622	567 (91.16%)	55 (0%)	2496	2435 (97.56%)	61 (0%)
Civil engr. and mechanical engr.	502	461 (91.83%)	41 (2.44%)	825	754 (91.39%)	71 (53.52%)
Electrical & electronic engr., information engr.	356	312 (87.64%)	44 (0%)	1730	1695 (97.98%)	35 (0%)
Chemical, materials, & Environmental engr.	734	630 (85.83%)	104 (0%)	2804	2738 (97.65%)	66 (0%)
Medical engr., biotechnology & Basic medicine	1121	1016 (90.63%)	105 (0.95%)	4433	4236 (95.56%)	197 (43.65%)
Clinical medicine I	893	833 (93.28%)	60 (5%)	5450	4633 (85.01%)	817 (90.7%)
Clinical medicine II & Dentistry	1196	1120 (93.65%)	76 (3.95%)	4988	4798 (96.19%)	190 (45.26%)
Health scis. and other medical scis.	863	751 (87.02%)	112 (10.71%)	3441	2679 (77.86%)	762 (78.61%)
Agricultural sciences	795	706 (88.81%)	89 (1.12%)	1987	1818 (91.49%)	169 (25.44%)
Economics and business	1242	904 (72.79%)	338 (0.89%)	2858	2244 (78.52%)	614 (21.82%)
(Interdisc.) social scis., media & Comm.	1575	1045 (66.35%)	530 (6.6%)	4008	1883 (46.98%)	2125 (74.35%)
Psychology and educational scis.	1230	893 (72.6%)	337 (4.75%)	3402	2054 (60.38%)	1348 (45.4%)
Political scis., public administration & Law	1080	590 (54.63%)	490 (4.69%)	2219	485 (21.86%)	1734 (73.41%)
Philosophy & theology	1144	588 (51.4%)	556 (3.78%)	1880	752 (40%)	1128 (41.13%)
Languages	974	401 (41.17%)	573 (1.92%)	1772	861 (48.59%)	911 (21.95%)
Literature, arts & architecture	1685	736 (43.68%)	949 (2.85%)	1622	535 (32.98%)	1087 (58.97%)
History, archaeology & cultural studies	1815	777 (42.81%)	1038 (5.01%)	4036	1102 (27.3%)	2934 (69.67%)
Multidisciplinary journals	34	26 (76.47%)	8 (50%)	1883	1656 (87.94%)	227 (87.67%)
Total	21,492	15,464 (71.95%)	6028 (3.6%)	67,608	52,569 (77.76)	15,039 (58.44%)

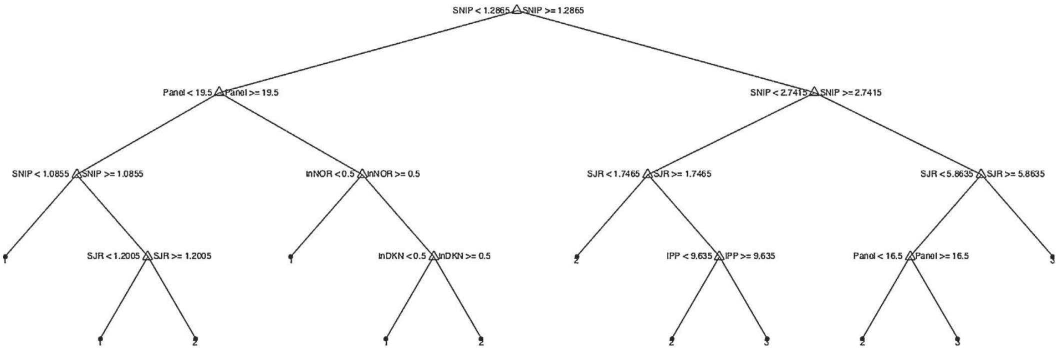


Fig. 2. Pruned decision tree.

Finnish journal publications (see last row in the table) are in journals that are covered in Scopus. However, for the journals assigned to the *Political Science and Public Administration* panel, less than 22% of the journal publications are in journals that are covered in Scopus. Similar observations have been made by Sivertsen (2014), who compared the coverage of journals and articles in the SSH to the other major disciplines from Norway's research institutions.

Table 8 also illustrates the language differences across disciplines. Altogether, the percentage of the not-covered-by-Scopus journals that are in Finnish and Swedish language is very small (see the fourth column in the table). However, for many disciplines (especially medicine and SSH) most of the not-covered-in-Scopus publications are articles in Finnish or Swedish language journals (see last column in the table). Again, this situation is comparable to that in Norway where only a few journals in the national language exist but a high percentage of the national articles from the SSH are concentrated in them (Sivertsen & Larsen, 2012). Actually, as described in (Puuska, 2014, pp.82–83), in both Norway and Finland groups of scholars and scientific societies had an effect of the higher rankings of publications with native languages.

4.2. Decision tree

Next, we built a decision tree for all 14,798 publication channels with the three indicators from Scopus (i.e. SJR, SNIP and IPP) available. 12,096 of these publication channels are rank 1, i.e. a trivial classifier predicting always 'rank 1' (compare Section 2.1) for this subset would be 81.74% correct. For our decision tree model, we used all variables that could have an effect on the expert-based rank decision in such a way that we utilized each variable from Table 3, either as it is if the variable had available values for all observations, or as a binary indicator on the availability of the variable if the variable had missing values. We used this strategy to ensure that the data set fed to the decision tree classifier had no missing values. Altogether, we had the three continuous variables (SJR, SNIP and IPP), two categorical variables (*panel* and *typeOfChannel*), and five binary variables (*inJCR*, *inNOR*, *inDNK*, *hasLanguageAvailable*, and *hasSherpaCodeAvailable*).

With stratified cross-validation (according to the four classes of ranks) and the inverse class frequencies as weights (see Section 3.1), we obtained a classifier that predicted the actual expert-based rank for nearly 88% of all publication channels correctly. Only 1853 (12.09%) of the publication channels were misclassified. In comparison with the trivial classifier, our decision tree was circa 6% more accurate. Fig. 2 shows the pruned decision tree. As can be seen from the figure, the SNIP indicator is the variable with the highest predictive power. However, also the other two Scopus metrics, as well as the panel and the information whether or not the publication channel is covered in the Norwegian and Danish databases, are important variables in the decision tree model.

4.2.1. Deviations

In Table 9, the characteristics of misclassifed observations are summarized, characterizing the subset of misclassified publication channels, for which (i) the Finnish expert-based rank was higher than the prediction (second column), and (ii) the subset for which the prediction was higher than the Finnish expert-based rank (third column) separately. For comparison reasons, the subset of correctly classified publication channels was characterized according to the same variables (fourth column in the table).

We see from Table 9 that the group of misclassified publication channels incorporates the most channels in which there has been a recent change in the expert-based rank. Interestingly, those misclassified publication channels with *Actual rank higher than prediction* have the highest percentage (12.5%) of positive change in rank, while those misclassified publication channels with *Actual rank lower than prediction* have the highest percentage (27.2%) of negative change in rank. The group with the highest percentage (86.6%) of publication channels in which the Finnish expert-based rank has not been changed recently indicates the correctly predicted observations. Moreover, we see from Table 9 that similar to the finding from the

Table 9
Characteristics of misclassified and correctly classified publication channels with decision tree.

	Misclassificed publication channels		Correctly classified publication channels
	Actual rank higher than prediction	Actual rank lower than prediction	
total sum	360 (2.43%)	1430 (9.66%)	13,008 (87.9%)
in JCR	154 (42.78%)	587 (41.05%)	7077 (54.4%)
mean number of publications	8.56	4.09	3.19
rank now higher	45 (12.5%)	9 (0.63%)	207 (1.59%)
no rankChange	286 (79.44%)	1032 (72.17%)	11265 (86.6%)
rank now lower	29 (8.06%)	389 (27.2%)	1536 (11.81%)
Language NaN	0 (0%)	0 (0%)	13 (0.1%)
English	309 (85.83%)	1245 (87.06%)	10865 (83.53%)
Finnish or Swedish	2 (0.56%)	1 (0.07%)	10 (0.08%)
other Language	49 (13.61%)	184 (12.87%)	2120 (16.3%)
Journal	360 (100%)	1430 (100%)	12999 (99.93%)
Conference	0 (0%)	0 (0%)	9 (0.07%)
NOR rank NaN	9 (2.5%)	43 (3.01%)	1501 (11.54%)
NOR rank equals FI rank	203 (56.39%)	1210 (84.62%)	10,146 (78%)
NOR rank higher FI rank	0 (0%)	78 (5.45%)	233 (1.79%)
NOR rank lower FI rank	148 (41.11%)	82 (5.73%)	934 (7.18%)
DNK rank NaN	8 (2.22%)	106 (7.41%)	2341 (18%)
DNK rank equals FI rank	237 (65.83%)	1070 (74.83%)	9481 (72.89%)
DNK rank higher FI rank	0 (0%)	194 (13.57%)	516 (3.97%)
DNK rank lower FI rank	115 (31.94%)	55 (3.85%)	669 (5.14%)

deviation study of association rules (Section 4.1.1), those publication channels for which the actual Finnish expert-based rank was higher than the prediction have with an average of 8.56, the most publications per publication channel.

4.3. Confusion matrix using reference indicator

As described in Section 3, we compared the *JuFo*-rank against the reference indicator using confusion matrices. As argued and concluded in Section 3.3.1, the three Scopus metrics met the requirement of fair external quality indicators the best. For interpretation purposes, we analyzed only that set of publication channels that had a highly deviating SJR, SNIP and IPP value (see Section 3.3).

To fractionalize the three Scopus metrics, we divided the available SJR, SNIP and IPP values into categories (0–3) such that the same frequencies of *JuFo*-ranks were present also in the SJR, SNIP and IPP categories. This fractionalization according to reference metrics was also used by Ahlgren and Waltman (2014) for the Norwegian expert-based ranking. With this rule, a publication channel is classified as rank 0 if the SJR value is smaller than 0.1, as rank 1 if SJR is in (0.1, 1.303], as rank 2 if SJR is in (1.303, 2.925], and, finally, as rank 3 if SJR is in (2.925, 45.894]. Similarly, SNIP is rank 1 if in (0, 1.442], rank 2 if in (1.442, 2.513, and rank 3 if in (2.513, 71.662]; and IPP is rank 1 if in (0, 2.419], rank 2 if in (2.419, 4.749], and rank 3 if in (4.7490, 159.283]. The confusion matrix between these sets are provided in Table 10.

We aim to have the same number of observations for each rank of the categorized Scopus metrics and the *JuFo*-ranking. However, the total number of observations of the *JuFo*-ranking and the categorized Scopus metrics do not coincide for each rank level. For example, 328 publication channels have an SJR value smaller than 0.1 and 444 publication channels have an SJR value of exactly 0.1. Therefore, we have 290 fewer publication channels that have SJR-rank 1 compared to the *JuFo*-rank (see *all* column and row for SJR in Table 10). Furthermore, 12,805 publication channels have an SJR value smaller than 1.303, and four publication channels have an SJR value of exactly 1.303, which results in three fewer publication channels that have SJR-rank 2 than *JuFo*-rank 2. Similar observations can be made for SNIP and IPP.

4.3.1. Deviations

As explained in Section 3, we entitle a publication channel to be highly deviating if the *JuFo*-rank is at least two ranks higher (or lower respectively) than the SJR, SNIP and IPP rank. As can be seen in Table 10, 225 publication channels are highly deviating in the sense that they have a higher, and 161 are highly deviating in the sense that they have a lower *JuFo*-rank than the SJR metric indicates. Furthermore, 259 publication channels have a higher *JuFo*-rank and 176 have a lower *JuFo*-rank than the SNIP metric indicates, while 427 publication channels have a higher *JuFo*-rank and 214 have a lower *JuFo*-rank than the IPP metric indicates. If we combine all subsets, 140 publication channels remain that have a higher *JuFo*-rank than all three reference indicators (the list of these publication channels can be found in Table A.16), and 60 have a lower *JuFo*-rank than all three indicators (these are explicitly listed in Table A.17).

Table 10
Confusion Matrices of fractionalized SJR, SNIP, IPP and *JuFo*.

	JuFo 0	JuFo 1	JuFo 2	JuFo 3	All
SJR 0	14	311	3	0	328
SJR 1	21	11080	1158	222	12,481
SJR 2	0	1219	646	191	2056
SJR 3	0	161	252	238	651
All	35	12771	2059	651	15,516
SNIP 0	1	657	108	4	770
SNIP 1	18	10257	994	147	11,416
SNIP 2	0	1076	710	263	2049
SNIP 3	0	176	238	237	651
All	19	12166	2050	651	14,886
IPP 0	0	654	108	4	766
IPP 1	12	9786	1234	315	11347
IPP 2	0	1445	444	151	2040
IPP 3	0	214	256	179	649
All	12	12,099	2042	649	14,802

Table 11 provides a summary of meta information for all publication channels for which the fractionalized SJR, SNIP and IPP are highly deviating. As can be seen from the table, the highly deviating channels combined make up less than 1% of all the publication channels in the system. Interestingly, we see again exactly as for the misclassified publication channels with decision tree in Section 4.2.1 that for the subset of publication channels in which the Finnish expert-based rank is higher than all three reference indicators (second column in the table), a high percentage of ranks has recently been changed to a higher rank, while for the publication channels for which the Finnish expert-based rank is lower than all three reference indicators suggest (third column in the table), a high percentage of ranks (70%) was recently changed to a lower one. Moreover, as already detected with the decision tree, we see that for the group for which the *JuFo*-rank is higher than suggested by all three reference indicators, on average, more publications of Finnish researchers exists. However, compared to the decision tree result, this time the difference is not significant.

The 140 publication channels that have a higher *JuFo*-rank than they should have according to the SJR, SNIP and IPP values can mostly be characterized by their SSH orientation (see Table A.16). This is, as discussed in Section 1, the underlying reason behind the expert-based final rankings according to the Norwegian model followed in Finland. As commented by Hicks (2012), SSH journals might be badly indexed in databases (like Scopus) and the language of the published articles can

Table 11
Characteristics of publication channels for which the expert-based rank is highly deviating from SJR, SNIP and IPP.

	Publication channels for which the *JuFo*-rank is highly deviating from SJR, SNIP and IPP	
	JuFo-rank higher than all three Scopus metrics	*JuFo*-rank lower than all three Scopus metrics
total sum	140(0.61%)	60(0.26%)
in JCR	9(6.43%)	53(88.33%)
mean number of publications	2.59	2.52
rank now higher	17(12.14%)	0(0%)
no rankChange	123(87.86%)	18(30%)
rank now lower	0(0%)	42(70%)
Language NaN	0(0%)	0(0%)
English	118(84.29%)	59(98.33%)
Finnish or Swedish	0(0%)	0(0%)
other Language	22(15.71%)	1(1.67%)
Journal	140(100%)	60(100%)
Conference	0(0%)	0(0%)
NOR rank NaN	0(0%)	6(10%)
NOR rank equals FI rank	125(89.29%)	45(75%)
NOR rank higher FI rank	0(0%)	7(11.67%)
NOR rank lower FI rank	15(10.71%)	0(0%)
DKN rank NaN	0(0%)	8(13.33%)
DKN rank equals FI rank	131(93.57%)	17(28.33%)
DKN rank higher FI rank	0(0%)	35(58.33%)
DKN rank lower FI rank	9(6.43%)	0(0%)

be other than English (see Table 8). This, with other disciplinary variations of the publication and citation patters (see Puuska (2014)), effects citation-based indicators. However, the SNIP indicator takes “subject field” into account (see Table A.13).

Moreover, for more than 93% of the 140 publication channels under study, the Danish expert-based rank (and for almost 90% of these publication channels, the Norwegian expert-based rank) coincides with the Finnish expert-based rank (see Table 11). If we combine all lists of publication channels, i.e. those which are evaluated higher in *JuFo* than the fractionalized SJR, SNIP and IPP, and those that have a Norwegian or Danish rank which is also lower than the *JuFo*-rank, only five publication channels remain.[7] Interestingly, again for four out of these five journal, the rank was recently updated to a higher one. Furthermore, according to *JuuliDB*, Finnish researcher have published in three of these journals, and in two cases the publications can again be linked to panel members.

The 60 publication channels that have a lower *JuFo*-rank than SJR, SNIP and IPP suggest can mostly be characterized by their review related orientation (see Table A.17). It is clear that review journals generally accumulate more citations than the original research articles. Therefore, they can be characterized by higher citation-based than expert-based rank. For 75% of these 60 publication channels the *JuFo*-rank coincides with the Norwegian (and for more than 28% with the Danish) expert-based rank (see Table 11). Altogether, only five publication channels were evaluated higher by the three citation-based reference indicators and the Norwegian and Danish expert-based ranks than by the Finnish experts.[8] Interestingly, again according to the pattern, the rank has recently been downgraded for all of these five channels. However, the most likely explanation why these seemingly very prestigious journals have not been ranked higher in *JuFo* is that all five, in fact, are review journals.

Summing up, for more than 99% of all publication channels under study (see the first row, *total sum*, in Table 11) the *JuFo*-rank was not highly deviating from SJR, SNIP and IPP. Moreover, for most of the publication channels for which the *JuFo*-rank was highly deviating from the three Scopus metrics, the *JuFo*-rank was supported by the Norwegian or Danish expert-based rank. Only ten publication channels (five for which the *JuFo*-rank was higher and five for which the *JuFo*-rank was lower) remained for which the *JuFo*-rank could not be explained by another citation- or expert-based metric.

5. Discussion and conclusions

The purpose of this study was to analyze whether or not the assignment of quality ranks to publication channels – which currently is performed by experts – could (at least partially) be replaced by automatic rules. We have provided an analysis of the national expert-based ranking system that used more variables, encodings, and computational methods than are found in the existing, relevant literature. Especially using novel techniques to cope with missing values (e.g., binary indication of whether a citation-based indicator is available or not) allowed us to analyze a much higher portion of the publication channels in *JuFoDB* than could have been analyzed by using other existing methodologies, which always restrict researchers to a subset of publication data and/or indicators that are completely available.

Association rules for the whole *JuFo* data showed that the availability of the three metrics provided by Scopus (SJR, SNIP and IPP), predict the Finnish expert-based rank very well. Furthermore, using decision trees with data for which the three Scopus measures were available, we found that a significant part of the work accomplished by the panels could be automated, or could at least provide a justified reference rank for panel discussion and decision-making. Similar to the study by Ahlgren and Waltman (2014), in which the Norwegian expert-based rank was predicted, our prediction model for the Finnish expert-based rank also showed that the SNIP indicator had the highest predictive power. The third part of our analysis illustrated that for more than 99% of the publication channels under study, the Finnish expert-based decisions did not deviate significantly from SJR, SNIP and IPP.

However, although the citation-based indicators showed the highest predictive power in our analysis, automatic rules using *only* these measures would certainly not be an alternative to the expert-based ranks. Ahlgren et al. (2012) concluded that with regard to coverage, currency, legitimacy, and transparency, the Norwegian model is preferable to automatic ranks constructed using citation-based indicators. Here, we argue that automatic rules could be utilized more under the condition that *all relevant and available* information is used to construct the prediction models. For example, our decision tree (Fig. 2) showed that besides the citation-based indicators, the panel (i.e., the discipline) of the publication channel – as well as the information whether the channel is covered in other relevant databases – are important variables to include in an automatic decision-support model. This fact was especially evident in Table 8, which showed the large disciplinary differences in coverage of both the journals in Scopus and the Finnish publications in them. Consequently, an automatic decision-support model should be based not only on citation-based indicators but also on information such as the discipline, language, and coverage in other databases.

Through our analysis of the publication channels for which the Finnish expert-based rank was higher than the rules suggested, we found multiple signs that the higher-than-predicted rank of a publication channel could be linked to the publication profile of Finnish scholars or even those who can influence the decision-making process. This discovery is

[7] *Etudes Classiques*, *Journal of Agricultural Science*, *Journal of American Folklore*, *Journal of Higher Education Policy and Management*, and *New German Critique*.

[8] *Biological Reviews*, *Natural Product Reports*, *Neuroscience and Biobehavioral Reviews*, *Progress in Neurobiology*, and *Trends in Plant Science*.

interesting when linked to the study by Serenko and Dohan (2011), who found a relationship between the current research interests of scholars and an overranking of publication channels in that particular research field. However, as also discussed by Ahlgren and Waltman (2014), opposite interpretations are possible for high deviations between expert-based ranks and citation-based indicators. Are these deviations a sign that expert-based opinions are truly necessary for avoiding the under- or overrating of certain publication channels? Or do they reveal (deliberate or unintentional) inaccuracies in the judgments of experts? We are not in the position to answer these questions, nor do we have the expertise to do so. However, our analysis of the highest deviating publication channels revealed certain patterns, and we think they should also be presented to the steering committee as part of the panel discussions.

In fact, interestingly, all three analysis methods showed that for the subset of publication channels with a higher-than-predicted rank, a high percentage of the Finnish expert-based rankings had been upgraded during the most recent panel evaluation. Similarly, in each case, the subset of publication channels with a lower-than-predicted rank showed the highest percentage of channels for which the rank had been downgraded during the most recent evaluation round. Basically, this result means that the old ranks coincided better with the other available quality information about the publication channels. However, as discussed in the paragraph above, there are two opposite interpretations of this finding that are possible.

As a whole, a data analysis methodology – expected ranks by a reference technique and the study of deviations – was proposed and demonstrated. This methodology can be applied in other similar instances of sparse data and tens of thousands of observations. From the report by Wilsdon et al. (2015), it is evident that automatization of expert judgment in research evaluation on the basis of advanced methodology and large datasets are currently a broad interest in research policy making. Naturally, this is possible only with open and accessible databases on publication channels and publication activity, according to the Norwegian model. Our analysis and results indicate that using repeatable methods and the detected rules and patterns, even if they are enlarged and improved (e.g. by considering also whether the publication channel publishes original research or reviews), could save money and man-hours in managing one of the three main components of the Norwegian performance-based funding model – the national database – and bring more transparency and objectivity into the second main component: the publication channel rankings.

Appendix A. Additional tables

Table A.12
Evaluating Panels: distribution of ranks.

Panel ID and name	Rank 0	Rank 1	Rank 2	Rank 3	Total
–	5329	1127	91	15	6562
1 Mathematics and statistics	5 (0.6%)	678 (82.1%)	106 (12.8%)	37 (4.5%)	826
2 Computer and Information Scis.	122 (8.2%)	1157 (78.0%)	153 (10.3%)	51 (3.4%)	1483
3 Physical Scis., Space Scis. & Astronomy	4 (0.8%)	472 (91.6%)	21 (4.1%)	18 (3.5%)	515
4 Chemical Scis.	1 (0.28%)	413 (92.6%)	25 (5.6%)	7 (1.6%)	446
5 Geosciences & Environmental Scis.	4 (0.68%)	571 (90.6%)	42 (6.7%)	13 (2.1%)	630
6 Biosciences I	5 (0.88%)	570 (91.5%)	36 (5.8%)	12 (2.0%)	623
7 Biosciences II	0 (0%)	568 (91.3183%)	34 (5.5%)	20 (3.2%)	622
8 Civil Engr. and Mechanical Engr.	25 (4.48%)	477 (84.9%)	48 (8.5%)	12 (2.1%)	562
9 Electrical & Electronic Engr., Information Engr.	32 (5.3%)	491 (81.0%)	68 (11.2%)	15 (2.5%)	606
10 Chemical, Materials, & Environmental Engr.	21 (2.8%)	669 (88.0%)	55 (7.2%)	15 (2.0%)	760
11 Medical Engr., Biotechnology & Basic Medicine	2 (0.2%)	1012 (89.8%)	92 (8.2%)	21 (1.9%)	1127
12 Clinical Medicine I	1 (0.1%)	815 (91.2%)	63 (7.1%)	15 (1.7%)	894
13 Clinical Medicine II & Dentistry	2 (0.2%)	1091 (90.9%)	81 (6.7%)	26 (2.2%)	1200
14 Health Scis. and other Medical Scis.	15 (1.7%)	781 (90.0%)	59 (6.8%)	13 (1.5%)	868
15 Agricultural sciences	2 (0.2%)	730 (91.5%)	51 (6.4%)	15 (1.9%)	798
16 Economics and Business	79 (6.2%)	1034 (81.7%)	117 (9.2%)	35 (2.8%)	1265

Table A.12 (*Continued*)

Panel ID and name	Rank 0	Rank 1	Rank 2	Rank 3	Total
17 (Interdisc.) Social Scis., Media & Comm.	26 (1.6%)	1317 (83.0%)	190 (12.0%)	55 (3.5%)	1588
18 Psychology and Educational Scis.	34 (2.7%)	1068 (84.4%)	120 (9.5%)	44 (3.5%)	1266
19 Political Scis., Public Administration & Law	11 (1.0%)	887 (82.0%)	147 (13.6%)	37 (3.4%)	1082
20 Philosophy & Theology	13 (1.1%)	959 (83.5%)	140 (12.2%)	36 (3.1%)	1148
21 Languages	23 (2.3%)	820 (81.0%)	138 (13.6%)	31 (3.1%)	1012
22 Literature, Arts & Architecture	11 (0.6%)	1413 (82.9%)	216 (12.7%)	65 (3.8%)	1705
23 History, Archaeology & Cultural Studies	11 (0.6%)	1507 (82.8%)	248 (13.6%)	53 (2.9%)	1819
24 Multidisciplinary journals	11 (30.6%)	10 (27.8%)	5 (13.9%)	10 (27.8%)	36
Total	5789 (20%)	20,637 (70%)	2346 (8%)	671 (2%)	29,443

Table A.13
Overview of reference quality indicators.

Indicator	Source	Original paper	Journal self-citations	Normalized	Description
Total Cites (ToCi)	Thomson Reuters' JCR	–	included	no	The total number of citations to the journal in the JCR year.
Impact Factor (IF)	Thomson Reuters' JCR	Garfield (1972)	included	yes	The average number of times articles from the journal published in the past two years have been cited in the JCR year. The IF is calculated by dividing the number of citations in the JCR year by the total number of articles published in the two previous years. An IF of 3.5 means that, on average, the articles published one or two year ago have been cited three and a half times. (Note that only citations that are indexed themselves in JCR contribute to the citation count.)
5-Year Impact Factor (5y IF)	Thomson Reuters' JCR	Fundamental idea goes back to Garfield (1972)	included	yes	The average number of times articles from the journal published in the past five years have been cited in the JCR year. It is calculated by dividing the number of citations in the JCR year by the total number of articles published in the five previous years.
Immediacy Index (II)	Thomson Reuters' JCR	Fundamental idea goes back to Garfield (1972)	included	yes	The average number of times an article is cited in the year it is published. The Immediacy Index is calculated by dividing the number of citations to articles published in a given year by the number of articles published in that year.
Articles	Thomson Reuters' JCR	–	not applicable	no	The total number of articles published in the journal in the JCR year.
Eigenfactor Score (EF)	Thomson Reuters' JCR	Bergstrom et al. (2008)	excluded	no	The Eigenfactor Score measures the importance of a citation by the influence of the citing journal divided by the total number of citations appearing in that journal. The calculation is based on the number of times articles from the journal published in the past five years have been cited in the JCR year, but it also considers which journals have contributed these citations so that highly cited journals will influence the network more than lesser cited journals.

Table A.13 (*Continued*)

Indicator	Source	Original paper	Journal self-citations	Normalized	Description
Article Influence Score (AI)	Thomson Reuters' JCR	Bergstrom et al. (2008)	excluded	yes	The journal's average EF score per published article. It is computed by dividing the EF through the number of articles published by the journal over the 5-year period.
IPP (Impact per Publication)	Scopus	Fundamental idea goes back to Garfield (1972)	included	yes	The impact per publication, calculated as the number of citations given in the present year to publications in the past three years divided by the total number of publications in the past three years.
SJR (SCImago Journal Rank)	Scopus	González-Pereira et al. (2010)	limited to max. one third	yes	The SJR is a measure of the scientific prestige of scholarly channels. SJR assigns relative scores to all of the channels in a citation network. Its methodology is inspired by the Google PageRank algorithm, in that not all citations are equal. A publication channel transfers its own 'prestige', or status, to another publication channel through the act of citing it. A citation from a publication channel with a relatively high SJR is worth more than a citation from a publication channel with a lower SJR. A publication channel's prestige for a particular year is shared equally over all the citations that it makes in that year; this is important because it corrects for the fact that typical citation counts vary widely between subject fields. The SJR of a publication channel in a field with a high likelihood of citing is shared over a lot of citations, so each citation is worth relatively little. The SJR of a publication channel in a field with a low likelihood of citing is shared over few citations, so each citation is worth relatively much. The result is to even out the differences in citation practice between subject fields, and facilitate direct comparisons of publication channels.
SNIP (Source Normalized Impact)	Scopus	Moed (2010)	included	yes	The SNIP per paper measures contextual citation impact by weighting citations based on the total number of citations in a subject field.

Table A.14
Availability of quality metrics in Thomson Reuters' JCR and their Spearman correlation to SJR, SNIP and IPP.

Thomson Reuters metric	Unavailable	Correlation		
	(isNaN)	SJR	SNIP	IPP
Total cites	0	0.3638	0.22226	0.34126
Impact factor	45	0.8709	0.80517	0.97411
5-Year impact factor	450	0.9076	0.76697	0.94966
Immediacy index	205	0.7500	0.66563	0.82632
Eigenfactor score	0	0.4123	0.23787	0.36926
Article influence score	450	0.9407	0.71489	0.86817

Table A.15
JuFo-rank 2 but no other reference of quality.

Name	Panel	SJR	IPP	SNIP
		Manually found		
Panel 2: 50				
ACM conference on computer and communications security	2	1.997	1.687	2.286
ACM conference on computer-supported cooperative work and social computing	2	coverage discontinued		
ACM international conference and exhibition on computer graphics and interactive techniques	2	–	–	–
ACM international conference on information and knowledge management	2	0.528	0.461	0.677
ACM international conference on mobile computing and networking	2	1.786	1.059	1.129
ACM international joint conference on pervasive and ubiquitous computing	2	coverage discontinued		
ACM multimedia conference	2	coverage discontinued		
ACM sigact-sigmod-sigart symposium on principles of database systems	2	2.208	1.554	1.518
ACM SIGCHI annual conference on human factors in computing systems	2	0.900	0.931	1.150
ACM sigkdd conference on knowledge discovery and data mining	2	2.879	2.023	2.331
ACM sigmod international conference on management of data	2	3.015	2.107	2.241
ACM sigplan conference on programming language design and implementation	2	3.141	2.099	2.768
ACM sigplan-sigact symposium on principles of programming languages	2	1.495	1.099	2.136
ACM sigsoft international symposium on the foundations of software engineering	2	coverage discontinued		
ACM symposium on computational geometry	2	0.670	0.548	0.770
ACM symposium on principles of distributed computing	2	1.127	0.894	1.165
ACM symposium on user interface software and technology	2	coverage discontinued		
ACM/IEEE international conference on human-robot interaction	2	–	–	–
ACM/siam symposium on discrete algorithms	2	2.247	1.520	1.644
Annual conference of the special interest group on data communication	2	–	–	–
Annual conference on neural information processing systems	2	–	–	–
Annual international acm sigir conference on research &development on information retrieval	2	coverage discontinued		
Computer aided verification	2	–	–	–
Conference on uncertainty in artificial intelligence	2	–	–	–
European conference on computer vision	2	–	–	–
European conference on information retrieval	2	–	–	–
European software engineering conference	2	–	–	–
European symposium on algorithms	2	–	–	–
IEEE annual symposium on foundations of computer science	2	–	–	–
IEEE international conference on data mining	2	–	–	–
IEEE international conference on pervasive computing and communications	2	coverage discontinued		
IEEE international symposium on parallel &distributed processing	2	coverage discontinued		
IEEE/ACM international conference on automated software engineering	2	–	–	–
International colloquium on automata, languages and programming	2	–	–	–
International conference on artificial intelligence and statistics	2	–	–	–
International conference on autonomous agents and multiagent systems	2	coverage discontinued		
International conference on information processing in sensor networks	2	coverage discontinued		
International conference on information systems	2	–	–	–
International conference on intelligent user interfaces	2	0.596	0.544	0.886
International conference on machine learning	2	coverage discontinued		
International conference on network protocols	2	–	–	–
International conference on pervasive computing	2	–	–	–
International conference on principles and practice of constraint programming	2	–	–	–
International conference on the theory and application of cryptographic techniques EUROCRYPT	2	–		

Table A.15 (*Continued*)

Name	Panel	SJR	IPP	SNIP
		Manually found		
International conference on tools and algorithms for the construction and analysis of systems	2	–	–	–
International cryptology conference CRYPTO	2	–	–	–
International semantic web conference	2	–	–	–
International symposium on software testing and analysis	2	–	–	–
Working IEEE/IFIP conference on software architecture	2	–	–	–
Www international conference on world wide web	2	–	–	–
Panel 5: 1				
Journal of geophysical research: oceans	5	2.031	3.108	1.249
Panel 17: 2				
Mir rossii	17	–	–	–
Yhteiskuntapolitiikka	17	–	–	–
Panel 18: 3				
Kasvatus	18	–	–	–
Language, cognition and neuroscience	18	0.0	0.0	0.0
Psykologia	18	–	–	–
Panel 19: 7				
Current legal problems	19	–	–	–
Hallinnon tutkimus	19	–	–	–
Legisprudence	19		coverage discontinued	
Mcgill law journal	19		coverage discontinued	
Oikeus	19	–	–	–
Politiikka	19	–	–	–
Zeitschrift fur europarecht, internationales privatrecht und rechtsvergleichung	19	–	–	–
Panel 22: 4				
Journal of dance education	22	–	–	–
Storyworlds	22	–	–	–
Taidehistoriallisia tutkimuksia	22	–	–	–
Theatre arts journal: studies in scenography and performance	22	–	–	–
Panel 23: 6				
Mitteilungen des Deutschen Archaologischen Instituts: Orient Abteilung: Baghdad	23		coverage discontinued	
Mitteilungen des Deutschen Archaologischen Instituts: Orient Abteilung: Damaskus	23	–	–	–
Studia fennica: anthropologica	23	–	–	–
Studia fennica: historica	23	–	–	–
Studia historica	23	–	–	–
Suomen muinaismuistoyhdistyksen aikakauskirja	23	–	–	–

Table A.16
JuFo-rank at least two ranks higher than SJR, SNIP and IPP.

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DAN
Panel 1: 1							
Journal of mathematical biology	1	3	1.183	1.432	2.017	2	2
Panel 2: 1							
Neural computation	2	3	0.878	1.13	1.572	2	2
Panel 15: 2							
Canadian journal of forest research-revue Canadienne de recherche forestiere	15	3	1.071	1.045	1.862	1	2
Journal of agricultural science	15	3	0.813	1.423	1.959	1	1
Panel 17: 4							
Acta sociologica	17	3	0.752	1.205	1.089	2	2
Communication monographs	17	3	1.024	1.223	1.326	2	2
Differences: a journal of feminist cultural studies	17	3	0.166	1.095	0.265	2	2
Feminist theory	17	3	0.672	1.299	0.782	2	2
Panel 18: 4							
Comparative education	18	3	0.812	0.766	0.747	2	2
Journal of cross-cultural psychology	18	3	0.917	1.228	1.64	2	2
Journal of higher education policy and management	18	3	0.881	1.151	1.008	1	1

Table A.16 (*Continued*)

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DAN
Journal of philosophy of education	18	3	0.687	1.221	0.622	2	2
Panel 19: 11							
Common market law review	19	3	0.645	1.12	0.495	2	2
Crime and delinquency	19	3	1.038	1.035	1.162	2	2
European journal of international law	19	3	0.573	1.36	0.616	2	2
European law journal	19	3	0.706	1.052	0.587	2	2
European law review	19	3	0.618	1.391	0.526	2	2
International and comparative law quarterly	19	3	0.59	1.156	0.484	2	2
Journal of law and society	19	3	0.381	1.16	0.656	2	2
Law and philosophy	19	3	0.352	1.026	0.269	2	2
Oxford journal of legal studies	19	3	0.454	1.121	0.554	2	2
Public management review	19	3	0.815	1.044	1.299	1	2
The modern law review	19	3	0.356	1.308	0.43	2	2
Panel 20: 19							
British journal for the history of science	20	3	0.254	1.391	0.462	2	2
Erkenntnis	20	3	0.621	0.961	0.49	2	2
International journal of systematic theology	20	3	0.139	0.257	0.033	2	2
Journal of biblical literature	20	3	0.332	0.564	0.106	2	2
Journal of contemporary religion	20	3	0.311	1.025	0.588	2	2
Journal of ecclesiastical history	20	3	0.166	0.443	0.106	2	2
Journal of the history of ideas	20	3	0.15	0.831	0.21	2	2
Journal of the history of philosophy	20	3	0.15	0.858	0.161	2	2
Method and theory in the study of religion	20	3	0.236	0.572	0.226	2	2
Neue zeitschrift fur systematische theologie und religionsphilosophie	20	3	0.104	0.075	0.031	2	2
New testament studies	20	3	0.353	1.089	0.224	2	2
Novum testamentum	20	3	0.123	0.282	0.036	2	2
Numen	20	3	0.133	1.015	0.175	2	2
Philosophy of science	20	3	1.086	1.303	0.877	2	2
Phronesis	20	3	0.159	1.44	0.231	2	2
Technology and culture	20	3	0.313	1.327	0.541	2	2
Vetus testamentum	20	3	0.196	0.131	0.014	2	2
Zeitschrift fur die alttestamentliche wissenschaft	20	3	0.28	0.049	0.01	2	2
Zeitschrift fur die neutestamentliche wissenschaft und die kunde der alteren kirche	20	3	0.121	0.209	0.043	2	2
Panel 21: 14							
Cognitive linguistics	21	3	0.718	1.356	0.812	2	2
English language and linguistics	21	3	0.544	1.254	0.686	2	2
Journal of African languages and linguistics	21	3	0.177	0.418	0.182	2	2
Journal of child language	21	3	1.04	1.329	1.475	2	2
Journal of pragmatics	21	3	1.038	1.226	0.909	2	2
Language variation and change	21	3	0.903	1.437	1.067	1	2
Langue francaise	21	3	0.331	1.055	0.222	2	2
Linguistic typology	21	3	0.304	0.445	0.385	2	2
Linguistics	21	3	0.584	1.068	0.642	1	2
Natural language engineering	21	3	0.316	1.126	0.695	2	2
Target: international journal of translation studies	21	3	0.293	1.323	0.39	2	2
Text and talk	21	3	0.433	0.656	0.339	2	2
Transactions of the philological society	21	3	0.339	1.159	0.351	2	2
Voprosy yazykoznaniya	21	3	0.1	0	0	2	2
Panel 22: 54							
Art history	22	3	0.13	0.505	0.067	2	2
Art journal	22	3	0.126	0.659	0.083	2	2
Boundary 2: an international journal of literature and culture	22	3	0.172	0.963	0.2	2	2
British journal of aesthetics	22	3	0.398	1.068	0.296	2	2
Burlington magazine	22	3	0.145	0.152	0.059	2	2
Cambridge opera journal	22	3	0.169	0.53	0.069	2	2
Cinema journal	22	3	0.138	0.916	0.25	2	2
Classical philology	22	3	0.132	0.463	0.05	2	2
Computer music journal	22	3	0.23	0.787	0.433	2	2
Critical quarterly	22	3	0.116	0.249	0.048	2	2
Design issues	22	3	0.274	0.832	0.571	2	2
Deutsche vierteljahrsschrift fr literaturwissenschaft und geistesgeschichte	22	3	0.115	0.098	0.029	2	2
Diacritics: a review of contemporary criticism	22	3	0.101	0	0	2	2
Early music history	22	3	0.169	0.743	0.167	2	2
Elh	22	3	0.148	0.879	0.102	2	2
Essays in criticism	22	3	0.1	0.15	0.025	2	2
Ethnomusicology	22	3	0.183	0.982	0.24	2	2

Table A.16 (*Continued*)

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DAN
Etudes classiques	22	3	0.1	0	0	1	1
History of photography	22	3	0.173	0.902	0.123	2	2
Journal of aesthetics and art criticism	22	3	0.203	0.514	0.161	2	2
Journal of architecture	22	3	0.242	0.372	0.143	2	2
Journal of design history	22	3	0.136	0.42	0.109	2	1
Journal of hellenic studies	22	3	0.179	0.549	0.065	2	2
Journal of musicology	22	3	0.208	0.839	0.105	2	2
Journal of new music research	22	3	0.26	0.857	0.575	2	2
Journal of the American musicological society	22	3	0.361	1.112	0.357	2	2
Journal of the royal musical association	22	3	0.233	0.691	0.184	2	2
Journal of the society of architectural historians	22	3	0.101	1.29	0.153	2	2
Journal of the warburg and courtauld institutes	22	3	0.111	0.895	0.086	2	2
Journal of visual culture	22	3	0.161	0.958	0.333	2	2
Leonardo	22	3	0.253	0.729	0.203	2	2
Mfs: modern fiction studies	22	3	0.179	0.578	0.094	2	2
Modern language quarterly	22	3	0.127	0.952	0.155	2	2
Music analysis	22	3	0.18	0.854	0.314	2	2
Music education research	22	3	0.761	1.075	0.519	2	2
Music theory spectrum	22	3	0.196	0.91	0.186	2	2
Narrative	22	3	0.181	1.138	0.23	2	2
New German critique	22	3	0.106	0.436	0.097	1	1
Nineteenth-century literature	22	3	0.127	0.574	0.111	2	2
October	22	3	0.1	0.508	0.049	2	1
Philologus	22	3	0.13	0.201	0.028	2	2
Popular music	22	3	0.201	1.188	0.348	2	2
Renaissance studies: journal of the society for renaissance studies	22	3	0.16	0.496	0.054	2	2
Representations	22	3	0.124	0.735	0.185	2	2
Screen	22	3	0.117	0.836	0.128	2	2
Scriptorium	22	3	0.1	0.557	0.043	2	2
Slavic and east European journal	22	3	0.126	0.353	0.104	2	2
Tdr	22	3	0.213	1.035	0.126	1	2
Television and new media	22	3	0.329	1.193	0.434	2	2
Textual practice	22	3	0.17	0.686	0.106	2	2
Theatre journal	22	3	0.192	0.985	0.143	2	1
Theatre research international	22	3	0.161	0.556	0.068	2	2
Yearbook for traditional music	22	3	0.168	0.839	0.188	1	2
Zeitschrift fur kunstgeschichte	22	3	0.1	0	0	2	2
Panel 23: 30							
American anthropologist	23	3	0.818	1.147	1.129	2	2
American antiquity	23	3	0.807	1.257	1.038	2	2
American journal of archaeology	23	3	0.376	1.284	0.422	2	2
Annales: histoire, sciences sociales	23	3	0.157	0.951	0.187	2	2
Anthropological theory	23	3	0.758	1.437	0.866	2	2
Antiquity	23	3	0.873	1.167	1.352	2	2
Archaeological dialogues	23	3	0.238	0.886	0.4	2	2
Early medieval Europe	23	3	0.137	1.291	0.261	2	2
Environmental history	23	3	0.28	0.987	0.44	1	2
Geschichte und gesellschaft	23	3	0.148	1.177	0.235	2	2
Historical methods	23	3	0.254	0.306	0.316	2	2
Historische zeitschrift	23	3	0.133	0.632	0.086	2	2
History	23	3	0.125	0.955	0.178	2	2
International history review	23	3	0.152	0.52	0.153	2	2
International review of social history	23	3	0.216	1.18	0.295	2	2
Jahrbucher fur geschichte osteuropas	23	3	0.155	0.91	0.109	1	2
Journal of American folklore	23	3	0.12	0.573	0.167	1	1
Journal of American history	23	3	0.16	1.125	0.267	2	2
Journal of contemporary history	23	3	0.186	1.159	0.33	2	2
Journal of folklore research	23	3	0.135	0.393	0.1	2	1
Journal of material culture	23	3	0.451	1.113	0.641	2	2
Journal of social history	23	3	0.165	0.882	0.212	2	2
Journal of womens history	23	3	0.257	1.389	0.449	2	2
Journal of world prehistory	23	3	0.829	0.885	1.724	2	2
Past and present	23	3	0.315	1.383	0.393	2	2
Rethinking history	23	3	0.227	1.156	0.349	2	2
Russian history	23	3	0.137	0.198	0.056	1	2
Scandinavian journal of history	23	3	0.153	1.041	0.274	1	2
Speculum: a journal of medieval studies	23	3	0.115	1.362	0.191	2	2
Vierteljahrshefte fur zeitgeschichte	23	3	0.145	1.425	0.27	2	2

Table A.17
JuFo-rank at least two ranks below SJR, SNIP and IPP.

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DKN
Panel 1: 1							
Archives of computational methods in engineering	1	1	6.284	5.712	7.175	1	1
Panel 2: 3							
ACM transactions on intelligent systems and technology	2	1	4.966	12.305	10.085	NaN	NaN
Foundations and trends in machine learning	2	1	12.076	17.015	19.5	NaN	1
Journal of statistical software	2	1	6.131	4.372	6.402	1	2
Panel 3: 11							
Advances in optics and photonics	3	1	7.988	9.249	11.28	NaN	NaN
Annual review of condensed matter physics	3	1	13.19	5.928	14.5	NaN	NaN
Astroparticle physics	3	1	3.012	2.776	3.828	1	1
Astrophysical journal letters	3	1	3.914	1.487	4.852	1	NaN
Astrophysical journal supplement series	3	1	6.857	3.125	9.687	2	1
Living reviews in solar physics	3	1	3.382	3.039	5.889	0	NaN
Monthly notices of the royal astronomical society	3	1	3.196	1.494	4.911	1	2
Monthly notices of the royal astronomical society: letters	3	1	3.661	1.503	4.106	NaN	NaN
Nano energy	3	1	3.403	2.379	5.951	NaN	NaN
Progress in quantum electronics	3	1	3.97	5.066	7.24	1	2
Publications of the astronomical society of the pacific	3	1	2.99	1.266	3.147	1	1
Panel 4: 13							
Accounts of chemical research	4	1	11.33	4.865	20.685	1	2
Acs catalysis	4	1	3.47	1.839	6.278	1	NaN
Acta crystallographica section D: biological crystallography	4	1	20.717	5.01	13.344	1	1
Aldrichimica acta	4	1	7.861	2.175	12.353	1	NaN
Annual review of analytical chemistry	4	1	3.082	2.445	7.841	NaN	NaN
Annual review of physical chemistry issn	4	1	7.602	4.836	14.741	1	1
Chemical society reviews	4	1	13.505	6.593	26.899	1	2
Coordination chemistry reviews	4	1	4.624	3.612	11.321	1	1
Journal of applied crystallography	4	1	3.119	6.457	5.829	1	2
Journal of photochemistry and photobiology c: photochemistry reviews	4	1	4.143	4.034	11.133	1	2
Mass spectrometry reviews	4	1	3.08	2.716	7.981	1	1
Natural product reports	4	1	3.116	3.778	9.338	2	2
Progress in solid state chemistry	4	1	3.448	6.624	7.692	1	1
Panel 5: 2							
Journal of the atmospheric sciences	5	1	3.464	1.491	2.992	2	1
Monthly weather review	5	1	4.039	1.692	3.345	1	1
Panel 6: 12							
Annual review of ecology evolution and systematics	6	1	6.226	4.259	13.275	1	2
Annual review of entomology	6	1	6.476	6.562	13.532	1	2
Annual review of phytopathology	6	1	6.037	4.47	12.257	1	2
Biological reviews	6	1	5.651	4.057	10.268	2	2
Current opinion in plant biology	6	1	5.656	2.201	8.833	2	2
Genome biology and evolution	6	1	3.162	1.017	4.314	1	NaN
Methods in ecology and evolution	6	1	2.946	2.384	4.64	1	NaN
Molecular ecology resources	6	1	3.468	2.927	6.913	1	1
Oceanography and marine biology	6	1	3.05	3.084	6	1	2
Quarterly review of biology	6	1	3.556	2.441	5.774	1	2
Studies in mycology	6	1	3.393	4.141	8.625	1	2
Trends in plant science	6	1	7.209	4.218	14.831	2	2
Panel 7: 43							
Advances in genetics	7	1	3.772	1.964	5.273	1	1
Annual review of biochemistry	7	1	27.902	6.978	29.52	1	2
Annual review of cell and developmental biology	7	1	19.686	4.777	20.105	1	2
Annual review of genetics	7	1	18.504	4.183	18.197	1	2
Annual review of microbiology	7	1	10.107	3.888	14.535	1	2
Biochimica et biophysica acta: gene regulatory mechanisms	7	1	3.642	1.309	5.607	1	1
Biochimica et biophysica acta: molecular cell research	7	1	2.999	1.344	4.93	1	1
Bioessays	7	1	3.251	1.139	4.577	2	1
Biotechnology advances	7	1	3.001	3.941	10.365	1	2
Cell reports	7	1	8.134	1.666	6.562	1	NaN
Chromosoma	7	1	2.942	0.756	3.068	1	1
Cold spring harbor perspectives in biology	7	1	4.857	1.276	4.689	1	NaN
Cold spring harbor symposia on quantitative biology	7	1	4.2	0.789	3.424	1	1
Critical reviews in biochemistry and molecular biology	7	1	5.107	1.558	6.436	1	2
Current opinion in biotechnology	7	1	3.382	2.146	7.812	1	2
Current opinion in cell biology	7	1	8.519	2.206	9.514	1	2
Current opinion in genetics and development	7	1	7.581	1.722	7.716	1	2

Table A.17 (*Continued*)

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DKN
Current opinion in microbiology	7	1	5.036	1.906	7.455	1	2
Current opinion in structural biology	7	1	6.88	1.987	8.377	1	2
Current opinion in virology	7	1	3.195	1.588	5.572	NaN	NaN
Current topics in developmental biology	7	1	3.988	1.385	5.457	1	1
Cytokine and growth factor reviews	7	1	3.939	2.488	9.133	1	2
Developmental biology	7	1	3.219	1.059	3.684	1	1
Epigenetics and chromatin	7	1	4.134	0.887	4.344	1	NaN
Fems microbiology reviews	7	1	7.649	4.143	13.299	1	2
Journal of biological chemistry	7	1	3.391	1.219	4.564	2	2
Journal of molecular biology	7	1	3.158	1.091	3.803	1	1
Journal of molecular cell biology	7	1	3.2	1.246	4.919	1	NaN
Microbiology and molecular biology reviews	7	1	10.607	5.107	16.429	1	2
Molecular plant	7	1	3.357	1.676	6.14	NaN	1
Mutation research: reviews in mutation research	7	1	3.285	2.041	6.719	1	2
Nature protocols	7	1	6.328	2.273	8.188	2	1
Nature reviews molecular cell biology	7	1	23.593	5.945	25.446	1	2
Open biology	7	1	4.545	1.25	4.23	NaN	NaN
Progress in lipid research	7	1	4.97	3.573	12.125	1	2
Reviews in medical virology	7	1	3.529	2.129	6.962	1	2
Seminars in cell and developmental biology	7	1	4.939	1.518	6.22	1	2
Trends in biochemical sciences	7	1	11.198	3.072	13.309	1	2
Trends in cell biology	7	1	10.198	2.71	11.754	1	2
Trends in genetics	7	1	9.354	2.263	10.754	1	2
Trends in microbiology	7	1	5.211	2.338	8.865	1	2
Wiley interdisciplinary reviews-computational molecular science	7	1	4.045	4.136	9.248	0	NaN
Wiley interdisciplinary reviews. rna	7	1	5.014	1.251	6.421	NaN	NaN
Panel 10: 2							
Annual review of chemical and biomolecular engineering	10	1	3.774	2.735	8.484	NaN	NaN
Geotechnique	10	1	3.91	3.156	2.372	1	2
Panel 11: 30							
Advances in immunology	11	1	4.303	1.447	5.271	1	2
Aids	11	1	3.701	1.756	5.759	2	1
Biochimica et biophysica acta: reviews on cancer	11	1	3.823	2.143	7.96	1	2
Brain behavior and immunity	11	1	2.967	1.447	5.83	2	1
Brain research reviews	11	1	4.54	2.903	8.682	1	2
Brain structure and function	11	1	3.304	0.942	3.365	2	1
Cancer discovery	11	1	4.676	1.13	5.129	1	NaN
Chemistry and biology	11	1	3.054	1.355	5.187	2	2
Circulation: cardiovascular genetics	11	1	3.337	1.35	5.563	1	NaN
Cold spring harbor perspectives in medicine	11	1	3.353	1.683	5.866	NaN	NaN
Current opinion in chemical biology	11	1	4.491	2.241	9.032	1	2
Current opinion in immunology	11	1	5.988	1.855	7.966	1	2
Current opinion in neurobiology	11	1	6.13	1.826	7.254	1	2
Developmental neurobiology	11	1	2.991	1.102	4.206	1	1
Disease models and mechanisms	11	1	3.06	1.308	4.856	1	NaN
Drug resistance updates	11	1	3.686	2.997	10.364	1	2
Frontiers in neuroendocrinology	11	1	3.632	2.34	8.49	2	2
Glia	11	1	3.15	1.41	5.452	1	2
Hippocampus	11	1	3.402	1.257	4.723	1	1
Immunological reviews	11	1	8.712	2.98	11.808	1	2
Molecular cancer therapeutics	11	1	3.117	1.441	5.926	1	1
Mucosal immunology	11	1	3.99	1.598	6.889	1	NaN
Neurobiology of disease	11	1	3.156	1.399	5.723	1	2
Neuroscience and biobehavioral reviews	11	1	5.666	3.344	10.596	2	2
Neuroscientist	11	1	3.392	1.891	7.075	1	1
Physiology	11	1	3.674	1.644	5.828	1	2
Progress in neurobiology	11	1	5.234	2.801	9.988	2	2
Seminars in immunology	11	1	4.207	1.081	5.262	1	2
Trends in immunology	11	1	7.5	2.412	10.435	1	2
Trends in neurosciences	11	1	10.184	3.393	13.309	1	2
Panel 12: 19							
Advances in cancer research	12	1	3.738	0.927	3.763	1	1
American heart journal	12	1	3.457	1.779	4.807	1	1
Cancer and metastasis reviews	12	1	4.053	2.157	8.21	1	2
Cancer treatment reviews	12	1	2.934	2.241	6.445	1	1
Circulation: arrhythmia and electrophysiology	12	1	3.968	2.081	5.288	1	NaN
Circulation: cardiovascular interventions	12	1	4.193	2.138	5.569	1	NaN
Circulation: cardiovascular quality and outcomes	12	1	4.515	2	4.989	1	NaN
Heart rhythm	12	1	3.335	1.744	4.209	1	1

Table A.17 (*Continued*)

Name	Panel	JuFo rank	SJR	SNIP	IPP	NOR	DKN
JAMA internal medicine	12	1	4.898	3.554	8.101	2	NaN
Journal of investigative dermatology symposium proceedings	12	1	4.223	3.412	7.875	1	1
Journal of mammary gland biology and neoplasia	12	1	3.22	1.783	6.596	1	1
Journal of thoracic oncology	12	1	3.051	1.87	5.394	1	1
Molecular oncology	12	1	3.5	1.392	5.926	1	NaN
Neoplasia	12	1	3.14	1.227	5.392	1	1
Neuro-oncology	12	1	3.023	1.741	6.012	1	NaN
Obesity reviews	12	1	3.638	2.904	8.497	1	2
Oncotarget	12	1	3.053	1.378	5.207	1	NaN
Seminars in cancer biology	12	1	5.117	2.108	8.265	1	2
Seminars in liver disease	12	1	3.471	2.855	8.045	1	1
Panel 13: 6							
Acta psychiatrica scandinavica	13	1	3.14	2.097	5.175	2	1
Alzheimers and dementia	13	1	5.814	4.251	13.075	1	NaN
Human reproduction update	13	1	4.341	4.107	9.89	1	1
Progress in retinal and eye research	13	1	5.174	4.087	10.778	1	2
Schizophrenia research	13	1	3.163	1.453	4.673	2	1
World psychiatry	13	1	3.34	4.073	7.074	1	2
Panel 14: 2							
Health affairs	14	1	4.636	3.001	4.538	1	2
Skeletal muscle	14	1	3.314	1.928	5.717	NaN	NaN
Panel 15: 1							
Renewable and sustainable energy reviews	15	1	3.273	3.644	6.822	1	1
Panel 16: 12							
Academy of management annals	16	1	9.928	4.74	8.225	NaN	1
Annual review of economics	16	1	7.843	2.514	2.849	1	NaN
Annual review of financial economics	16	1	3.706	1.447	1.426	1	NaN
Brookings papers on economic activity	16	1	5.301	2.708	2.308	1	1
Computers and operations research	16	1	2.97	2.942	3.076	2	1
Economic policy	16	1	5.212	4.003	3.875	1	1
Imf economic review	16	1	4.335	2.602	2.563	1	1
Journal of economic perspectives	16	1	8.485	5.176	5.138	1	2
Nber macroeconomics annual	16	1	3.03	1.498	1.182	1	1
Qme: quantitative marketing and economics	16	1	3.976	0.897	1.238	1	2
Review of environmental economics and policy	16	1	3.175	2.58	3.661	1	1
Tax policy and the economy	16	1	3.22	3.012	2.125	1	1
Panel 17: 1							
Foundations and trends in communications and information theory	17	1	6.471	3.324	4.778	1	1
Panel 18: 3							
Neuropsychology review	18	1	3.193	2.432	6.861	1	1
Perspectives on psychological science	18	1	5.179	3.892	7.596	1	1
Psychological science in the public interest	18	1	4.451	9.167	12.75	1	1

References

Abramo, G., & D'Angelo, C. A. (2015). Evaluating university research: Same performance indicator, different rankings. *Journal of Informetrics*, *9*(3), 514–525. http://www.sciencedirect.com/science/article/pii/S1751157715000462

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record. Vol. 22* (pp. 207–216). ACM. http://www.almaden.ibm.com/cs/quest/papers/sigmod93.pdf

Ahlgren, P., Colliander, C., & Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics*, *92*(3), 767–780. http://dx.doi.org/10.1007/s11192-012-0632-x

Ahlgren, P., & Waltman, L. (2014). The correlation between citation-based and expert-based assessments of publication channels: SNIP and SJR vs. Norwegian quality assessments. *Journal of Informetrics*, *8*(4), 985–996. http://www.sciencedirect.com/science/article/pii/S1751157714000911

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). Cambridge, MA, USA: The MIT Press. https://mitpress.mit.edu/books/introduction-machine-learning

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, *79*(3), 635–649. http://dx.doi.org/10.1007/s11192-007-2036-x

Auranen, O., & Nieminen, M. (2010). University research funding and publication performance – An international comparison. *Research Policy*, *39*(6), 822–834. http://www.sciencedirect.com/science/article/pii/S0048733310000764

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *Journal of Neuroscience*, *28*(45), 11433–11434. http://www.jneurosci.org/content/28/45/11433.short

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, *4*(6), e6022. http://dx.doi.org/10.1371/journal.pone.0006022

Bornmann, L., Marx, W., Schier, H., Rahm, E., Thor, A., & Daniel, H.-D. (2009). Convergent validity of bibliometric Google Scholar data in the field of chemistry – Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics*, *3*(1), 27–35. http://www.sciencedirect.com/science/article/pii/S1751157708000667

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bryman, A. (2004). Triangulation. In *The SAGE encyclopedia of social science research methods* (pp. 1143–1144). Sage Publications, Inc. http://dx.doi.org/10.4135/9781412950589

Butler, L. (2003). Explaining Australias increased share of {ISI} publications the effects of a funding formula based on publication counts. *Research Policy*, *32*(1), 143–155. http://www.sciencedirect.com/science/article/pii/S0048733302000070

Cattaneo, M., Meoli, M., & Signori, A. (2014). Performance-based funding and university research productivity: The moderating effect of university legitimacy. *Journal of Technology Transfer*, *41*(1), 85–104. http://dx.doi.org/10.1007/s10961-014-9379-2

Chang, C.-L., McAleer, M., & Oxley, L. (2013). Coercive journal self citations, impact factor. *Journal Influence and Article Influence, Mathematics and Computers in Simulation*, *93*, 190–197. http://www.sciencedirect.com/science/article/pii/S0378475413000694

Danish Centre for Studies in Research and Research Policy. (2014). *Evaluation of the Norwegian publication indicator – English summary. Tech. rep.* Danish Centre for Studies in Research and Research Policy. http://www.uhr.no/documents/Evaluation_of_the_Norwegian_Publication_Indicator___English_Summary.pdf

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics*, *9*(4), 895–906. http://www.sciencedirect.com/science/article/pii/S1751157715300894

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *FASEB journal*, *22*(8), 2623–2628. http://www.ncbi.nlm.nih.gov/pubmed/18408168

Franceschet, M. (2010). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, *4*(1), 55–63. http://www.sciencedirect.com/science/article/pii/S1751157709000698

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. In *Science* (pp. 471–479). American Association for the Advancement of Science. http://www.elshami.com/Terms/I/impact%20factor-Garfield.pdf

Giovanni, A., Tindaro, C., & D'Angelo, C. A. (2014). Are the authors of highly cited articles also the most productive ones? *Journal of Informetrics*, *8*(1), 89–97. http://www.sciencedirect.com/science/article/pii/S1751157713000886

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391. http://www.sciencedirect.com/science/article/pii/S1751157710000246

Good, B., Vermeulen, N., Tiefenthaler, B., & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. *Research Evaluation*, *24*(2), 91–105.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, *6*(4), 674–688. http://www.sciencedirect.com/science/article/pii/S1751157712000521

Haustein, S., & Larivière, V. (2015). The use of bibliometrics for assessing research: possibilities, limitations and adverse effects. In *Incentives and performance* (pp. 121–139). Springer. http://link.springer.com/chapter/10.1007%2F978-3-319-09785-5_8

He, H., Garcia, E., et al. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. http://dx.doi.org/10.1109/TKDE.2008.239

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. http://www.sciencedirect.com/science/article/pii/S0048733311001752

Kärkkäinen, T., & Saarela, M. (2015). Robust principal component analysis of data with missing values. In *Machine learning and data mining in pattern recognition* (pp. 140–159). Springer. http://link.springer.com/chapter/10.1007%2F978-3-319-21024-7_10

Kreiman, G., & Maunsell, J. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience*, *5*, 48. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3214728/

Kuroda, M., Mori, Y., Masaya, I., & Sakakihara, M. (2013). Alternating least squares in nonlinear principal components. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 456–464. http://dx.doi.org/10.1002/wics.1279

Moed, H. F. (2005). *Citation analysis in research evaluation*. The Netherlands: Springer. http://www.springer.com/us/book/9781402037139

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. http://www.sciencedirect.com/science/article/pii/S1751157710000039

Pruvot, E. B., Claeys-Kulik, A.-L., & Estermann, T. (2015). Strategies for efficient funding of universities in Europe. In A. Curaj, L. Matei, R. Pricopie, J. Salmi, & P. Scott (Eds.), *The European higher education area: Between critical reflections and future policies* (pp. 153–168). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-20877-0_11

Puuska, H.-M. (2014). *Scholarly publishing patterns in Finland – A comparison of disciplinary groups, (Ph.D. thesis)*. Tampere University Press. https://tampub.uta.fi/bitstream/handle/10024/95381/978-951-44-9480-2.pdf?sequence=1

Saarela, M., & Kärkkä inen, T. (2015). Analysing student performance using sparse data of core bachelor courses. *JEDM-Journal of Educational Data Mining*, *7*(1), 3–32. http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM056

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science*, *8*(3), 364–378. http://dx.doi.org/10.1057/eps.2009.19

Schneider, J. W., Aagaard, K., & Bloch, C. W. (2015). What happens when national research funding is linked to differentiated publication counts? A comparison of the Australian and Norwegian publication-based funding models. *Research Evaluation*, 1–13.

Seiler, C., & Wohlrabe, K. (2014). How robust are journal rankings based on the impact factor? Evidence from the economic sciences. *Journal of Informetrics*, *8*(4), 904–911. http://www.sciencedirect.com/science/article/pii/S1751157714000820

Serenko, A., & Dohan, M. (2011). Comparing the expert survey and citation impact journal ranking methods: Example from the field of artificial intelligence. *Journal of Informetrics*, *5*(4), 629–648. http://www.sciencedirect.com/science/article/pii/S1751157711000666

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter*, *6*(1), 22–28.

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in scopus and web of science. In E. Noyons (Ed.), *Proceedings of the science and technology indicators conference* (pp. 598–604). Leiden: Universiteit Leiden.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics*, *91*(2), 567–575. http://dx.doi.org/10.1007/s11192-011-0615-3

Vanclay, J. K. (2011). An evaluation of the Australian Research Council's journal ranking. *Journal of Informetrics*, *5*(2), 265–274. http://www.sciencedirect.com/science/article/pii/S1751157710000994

Vanclay, J. K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, *32*(2), 211–238. http://dx.doi.org/10.1007/s11192-011-0561-0

Vanclay, J. K., & Bornmann, L. (2012). Metrics to evaluate research performance in academic institutions: a critique of ERA 2010 as applied in forestry and the indirect H_2 index as a possible alternative. *Scientometrics*, *91*(3), 751–771.

Verleysen, F. T., Ghesquière, P., & Engels, T. (2014). The objectives, design and selection process of the Flemish academic bibliographic database for the social sciences and humanities (vabb-shw). In *Bibliometrics: Use and abuse in the review of research performance*. pp. 117–127.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Hill, J., Kain, R., Johnson, B., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. Tech. rep.* Higher Education Funding Council for England. http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html