Anu Koskela

Exploring the differences of Finnish students in PISA 2003 and 2012 using educational data mining

Master's Thesis in Information Technology

November 18, 2016

University of Jyväskylä

Department of Mathematical Information Technology

Author: Anu Koskela

Contact information: anu.k.koskela@student.jyu.fi

Supervisors: Tommi Kärkkäinen, and Mirka Saarela

Title: Exploring the differences of Finnish students in PISA 2003 and 2012 using educational data mining

Työn nimi: Erojen tutkiminen vuosien 2003 ja 2012 suomalaisten oppilaiden PISA-tulosten välillä tiedonlouhintaa käyttäen

Project: Master's Thesis

Study line: Computational Science

Page count: 40+2

Abstract: Finland has always gotten good scores in PISA studies, but in 2012 the results dropped significantly. A modified version of a data mining algorithm called k-means++ was applied to 2003 and 2012 Finnish PISA data. As a result five clusters were created for both data sets and the results were compared. The clusters with lower results had remained similar but the cluster with higher scores had been divided. Additionally the medium cluster had grown significantly. The results in mathematics had dropped in all clusters since 2003.

Keywords: Master's Thesis, Data Mining, K-means++, Clustering, PISA

Suomenkielinen tiivistelmä: Suomi on aina saanut hyviä tuloksia PISA-tutkimuksissa, mutta vuonna 2012 tulokset huononivat merkittävästi. Vuosien 2003 ja 2012 aineistoihin sovellettiin muokattua tiedonlouhinta-algoritmia nimeltään k-means++. Tuloksena saatiin kummallekin aineistolle viisi klusteria, joita vertailtiin keskenään. Huonommin pärjänneet klusterit olivat pysyneet samankaltaisina, mutta paremmin pärjänneiden klusteri oli jakautunut. Lisäksi keskiverrosti pärjänneiden klusteri oli kasvanut huomattavasti. Kaikkien klustereiden tulokset matematiikassa olivat laskeneet vuodesta 2003.

Avainsanat: pro gradu -tutkielma, tiedonlouhinta, k-means++, klusterointi, PISA

Glossary

Centroid	A center point in a cluster (see ”cluster”) which represents the points belonging to that cluster
Cluster	Group of observed points that resemble each other
Educational Data Mining	Using data mining techniques to education-related data
KDD	Knowledge Discovery in Databases, a process model
K-means	A data-mining method used to find clusters in large data sets
OECD	Organisation for Economic Co-operation and Development
PISA	Program for International Student Assessment
Ray-Turi index	A calculated index describing the quality of a data mining result
Scale index	A variable which is created by combining multiple of the original variables used in a study to represent a certain feature or quality

List of Figures

List of Tables

Contents

1 Introduction

Large-scale student assessment has been very popular in the USA especially after the 1970's (Hoepfl 2000), and the amount of testing has been increasing ever since (Shavelson, Schneider, and Shulman 2007). Traditionally it has not been used that much in Europe (De Corte, Janssen, and Verschaffel 2009). The aim of student assessment is to get more information about the students' abilities, the curriculum, and the teaching, in order to improve the curriculum and teacher training and thus the learning results. The tests may focus on a certain subject or area of a curriculum, and may include questions about the students' background to provide more explanation for the results. In large-scale assessment there are often thousands of students participating, and the participants can represent many different schools and regions. It has been argued that the tests may not actually improve teaching in large scale, but may lead to teachers focusing more on what the tests are about, and thus narrowing the fields covered in teaching. This phenomenon is known as ”What You Test Is What You Get” (De Corte, Janssen, and Verschaffel 2009). Regardless, the demand for testing seems to remain high.

In 2000 the Organisation for Economic Co-operation and Development (OECD) organized the Programme for International Student Assessment (PISA) for the first time. PISA focuses on assessing the learning and skills of roughly 15-year-old students all over the world. It is organized triennially and has become well known, attracting more participating countries each time. Each time the PISA studies are conducted, they focus on one of the main topic: mathematics, reading or science.(*”About PISA”* n.d.) Finland has received top results in PISA tests, but in the most recent studies the results of Finnish students have dropped both compared to other countries and to the earlier test results of Finnish students. This thesis aims to find out if there are clear differences between the mathematics results and the student profiles between the 2003 and 2012 PISA studies.

Chapter 2 introduces the PISA study, gives more information about how the tests are designed, structured and conducted, and presents some of the results of Finnish students. Chapter 3 explains the concept of data mining and introduces the algorithm which is used in this thesis to analyze the data. The data and the process of the conducted study is described in chapter 4, and the results are presented in chapter 5 along with some analysis on the results. There is a conclusion at the end, describing shortly the data, the process and the obtained results, and providing some ideas for further studies.

2 Programme for International Student Assessment

The Organisation for Economic Co-operation and Development (OECD) has a Programme for International Student Assessment (PISA) which tests 15-year-old students' knowledge and learning outcomes in different countries every three years (Saarela and Kärkkäinen 2014), (Economic Co-operation and Development 2003). Each time the PISA test is conducted, there is a focus on either mathematics, reading or science (*"About PISA"* n.d.). The PISA test was first conducted in 2000. There were 32 countries participating and between 4500 and 10000 students from each country (Economic Co-operation and Development 2003). In 2012 there were 65 countries and about 510 000 students participating (PISA 2012).

2.1 Test planning and conduction

From each participating country a subset of schools takes part in the PISA study. The schools are selected so that they can represent the whole population of the country. There are separate variables for assessing the schools and different for the students. The participating students answer questions about their economic, social and cultural background. They also answer questions that are related to their motivation, tendencies, skills and learning outcomes related to different subjects. In addition, there are tasks which the students will solve, for example mathematical problems. Each student has about 30 minutes for answering the background questionnaire and two hours for completing the presented tasks.

Originally each students was given the full student questionnaire to answer. In 2012 a planned rotation of questions was used; The questions were divided into three groups and the majority of participating students were given two of the three sets of questions. Only a minority of students were asked to answer all three question sets. Certain background questions, such as gender, were presented to all students. The planned rotation means that roughly a third of the data is missing on purpose. Since there were so many participating countries and students in 2012, the amount of missing data was not an issue: there is still plenty of data, and the rotation enabled having more questions and thus getting more diverse data for the analysis. (OECD 2014, page 58)

In PISA data, as in most other studies, the final data set represents a larger population than what took part in the study. For example, in 2012 there were 8829 students participating in Finland, but the study represents the entire population of 60047 15-year-old Finnish students at the time (Saarela and Kärkkäinen 2015). In 2012, immigrants and Swedish-speaking students were overrepresented in the PISA sample on purpose, which means that the data had to be somehow scaled so that it could represent the whole population. This is why each student was given a weight, which can be found in both 2003 and 2012 data as a variable: student final weight (FSTUWT) (OECD 2005) and (OECD 2014). Weight of 1 means that the student represents only himself/herself, and the higher the weight, the larger number of students the individual represents. Figure 1 presents the density of different student weights in both 2003 and 2012 PISA data.

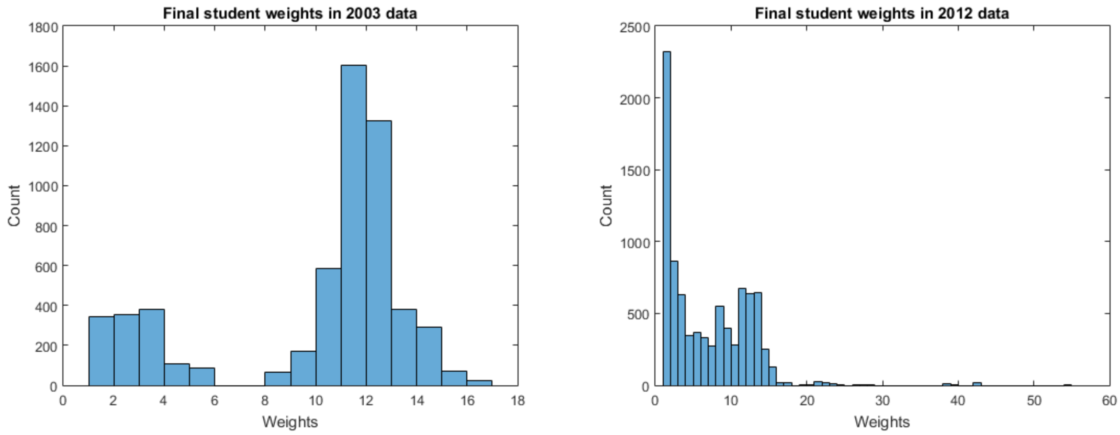

Figure 1. Final student weights

One way to ease the comparison of the results was to calculate so called *plausible values* (PVs). First for each question a so called difficulty level was calculated based on how many students had solved that task, and for each student a competence value was calculated based on how many tasks the student had solved. Based on these numbers PVs were created for each domain (mathematics, reading and science). The PVs cannot be directly connected to an individual student, but they represent characteristics of the population.(OECD 2014) Each time a domain was in focus, the PVs were calculated for also the subdomains.(Saarela and Kärkkäinen 2014)

2.2 Finnish results

In the earliest PISA studies Finland received spectacular results. In the first PISA test in 2000 Finland got top scores, and ranked first out of all participating countries in reading literacy, the third in science literacy and fourth in mathematical literacy (the main focus of the study that year was reading literacy)(Education and Culture, n.d.[b]). In 2003 Finland was first in reading and science literacy and second in mathematical literacy and problem solving(Education and Culture, n.d.[c]). 2009 was the first time when Finland did not rank first in any of the categories when looking at all of the participating countries, but even then Finland ranked first in scientific literacy within OECD countries(Education and Culture, n.d.[d]).

Finland's excellent results were unexpected in the first PISA studies and as a result Finland's education system gained a reputation as being the best in the world. There are several publications that analyze the Finnish school system and try to figure out what it is that makes it work, including (Välijärvi et al. 2007), (Kupiainen, Hautamäki, and Karjalainen 2009) and (Education and Culture, n.d.[a]). The following paragraphs presents some overview on the topic.

It has been speculated what is the reason behind Finland's success in the PISA studies. One reason might be the school reform that was done in the 1970's. The main goal then was to create a system where each child could have an equal opportunity to get a good education, and it is still in today's legislation (see *"Basic Education Act 628/1998"* n.d. Section 2). There are no tuition fees, all students receive free books, notebooks, pencils etc. There is also free health care and a free lunch at school every day. There are only very few private schools and the public schools generally have a very good quality teaching. There are no national tests during primary and secondary school, and only some guidelines for what the students should learn, leaving teachers freedom to plan the teaching and testing. There are hardly any special classes based on the students' performance. This means that each class has both well and low performing students, and the low performing students are given extra help when it is needed. As a result, even the low performing students are performing decently related to other countries, and this as a whole raises Finland's results.

It is said that the main reasons why this type of system can function well in Finland, is that the population is very homogenous and small, and it is scattered on a vast area geographically. The Finnish culture also values education, and teachers are generally respected. Teachers are required to have a master's degree, and only roughly 10 percent of the applicants get to study pedagogy. The aim of teaching is generally to help students to understand different concepts and try to find different solutions to problems rather than simply teaching rules and formulas on how to solve specific problems. (Education and Culture, n.d.[a])

In the most recent results (PISA study of 2012) the performance of Finnish students has decreased. In the PISA test of 2012 Finland ranked fifth in scientific literacy, sixth in reading literacy and 12th in mathematical literacy when looking at all participants(Education and Culture, n.d.[e]). The results have dropped not only internationally but also nationally compared to the earlier results. The reasons why the performance has gotten lower is not clear, but it is clearly an important issue.(*"PISA 12 Still among the best in the OECD but performance is declining"* n.d.)

3 Data Mining

Data mining uses various algorithms to handle large data sets. Usually the aim is to find patterns and new models from large-scale data. This provides more information about the data and can help to describe, understand and predict the models. The difference between statistical and data mining methods is that data mining can use not only statistics but also machine learning, database systems and pattern recognition (Zaki and Meira Jr 2014), thus discovering phenomena that might otherwise be missed. Data mining can be used widely in different fields. For example, educational data mining is using data mining methods to analyse data that can be related to various aspects of teaching and learning.

It is important to understand the characteristics of the data being used in the research. For example: are there a lot of missing values, are they missing at random or is there a pattern? Does the data represent the whole population that is the target of the study and has the sampling been done fully random or are some groups missing or overrepresented? Is the data numeric, textual or something else? The structure of the data affects the selection of possible methods that may be applied. In many cases the data may be pre-processed to be more suited for a specific method. For example, the missing values may be replaced or the rows with missing values excluded.

Data mining techniques can be categorized based on what the approach and the goal is. In *exploratory data analysis*, the goal is to find ”the key characteristics of the data sample via statistics [-]” (Zaki and Meira Jr 2014). *Frequent pattern mining* methods aim to discover patterns and trends in the data to understand how different entities and variables relate to each other. In *clustering* algorithms the data is divided into natural groups where the entities within one group are similar to each other and dissimilar to entities in other groups. In *classification* the ”task is to predict the label or class for a given unlabeled point.”(Zaki and Meira Jr 2014) Often different methods can be combined; One method can be used to create a guess or a rough estimate of the solution, and another method can then be applied by using the results from the previous method as starting points which will then be improved. Other methods can also be used in the analysis and validation of the gained results.

One area of data mining is clustering and one of the basic clustering algorithms is called *k-means*. The following description of the basic k-means algorithm is based on book (Zaki and Meira Jr 2014). The k-means algorithm divides the data into k clusters and the value of k has to be set in advance. The purpose is to divide the data so that the points in each cluster are similar to each other and dissimilar to the points in other clusters. The higher the amount (k) of clusters, the harder it can be to find such clusters and an estimation has to be taken into consideration. This method iteratively increases the accuracy of the clusters.

The basic idea of the algorithm is as follows. First, choose k points within the data set. The points can be observed points or calculated values (such as mean), depending on the variant of the algorithm, and they can be selected by random or by using some other algorithm. The second step is to calculate the distances (can be Euclidean or other way of measuring distances) between each observation point and each originally selected k points. The original k points are called centroids or prototypes, and the clusters are constructed by assigning each point to the nearest centroid. The third step is to calculate new centroids for each cluster, for example by calculating the mean of each cluster.

The algorithm repeats steps two and three until no observed points change cluster, or until the maximum amount of iterations is reached. For example in Matlab the k-means algorithm uses the Euclidean distance and the default number of iterations is 100. K-means is a greedy method and may end up finding a local minimun instead of a global one, depending on the starting points. The likelihood of finding a global minimum can be increased by running the algorithm multiple times and selecting the best result.

Since the result may depend on the selection of the initial points, it makes sense to improve the selection of those points. In *k-means++* one point of the group of points is selected (uniformly at random) to be the first center point. Then calculate the distance between each of the other points to the selected centroid and select a new centroid at random using a weighted probability distribution (proportional to the squared distance of that point to the previous centroid). Continue selecting new centroids this way until k centroids have been selected. Then proceed with the original k-means. It has been shown that using k-means++ improves the results and reduces the clustering time. (Arthur and Vassilvitskii 2007)

The basic k-means and k-means++ algorithms do not take weights into consideration. It was discovered in (Saarela and Kärkkäinen 2015) that adding weights to both the initialization and clustering phases in k-means++ reduced both the required amount of steps and the clustering error in the next phase. Since it is known that the data used in this thesis is biased by design (see section 2.1), the weights needed to be included in the clustering. The full k-means++ with weights algorithm implementation (which was also used in (Saarela and Kärkkäinen 2015)) can be found in appendix A, listing 6.1.

4 KDD Process

The process leading to and including data mining and analysis follows the knowledge discovery in databases (KDD) process which is described in (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Figure 2 displays the steps included in the process. The steps in this thesis followed the KDD process, and the steps are described in the following sections: section 4.1 is the selection step, section 4.2 is about preprocessing and transformation and section 4.3 is related to the data mining. The evaluation and interpretation steps can be found in chapter 5.

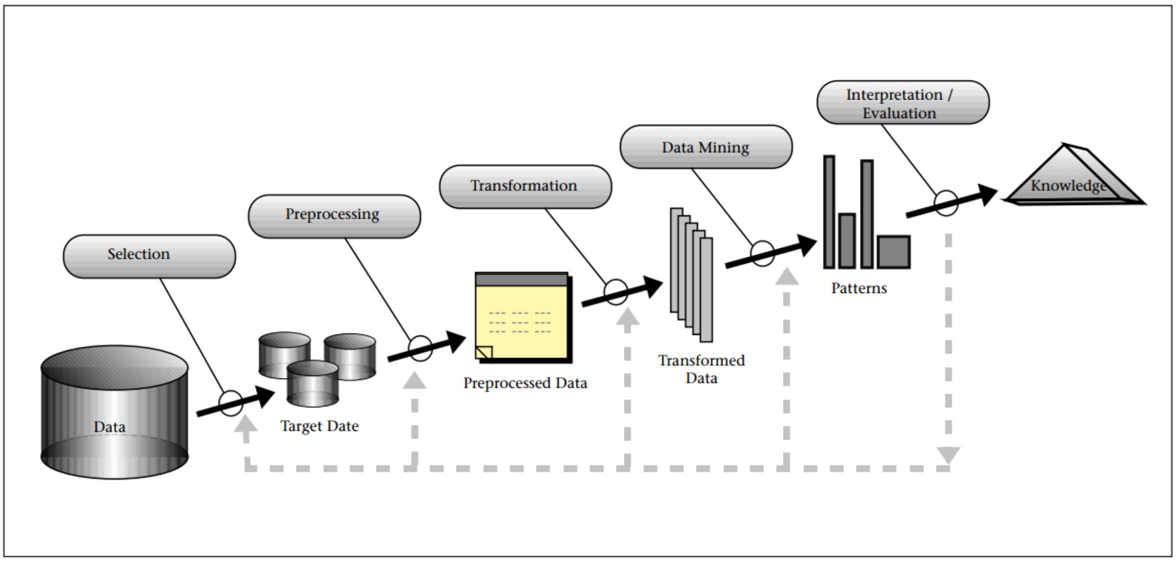

Figure 2. KDD process according to (Fayyad, Piatetsky-Shapiro, and Smyth 1996)

4.1 Scale indices

All of the questions and data from PISA studies can be downloaded from https://www.oecd.org/pisa/pisaproducts/. For this thesis full student data was downloaded for 2003 and 2012. The data comes as .spss data format and needs to be transformed before it can be fully used in Matlab. Basically this means opening the data in SPSS and saving it into .dat format. Once the data had been saved in the correct format, it was downloaded into Matlab by using the *readtable* command. The subset of only Finnish data was saved into a separate data table to be used, and the rest of the data was discarded.

Since the data includes a massive amount of variables, some decisions had to be made in order to narrow down the variable set. There are so called scale indices in PISA data which represent certain properties by combining multiple variables. Scale indices have also been scaled so that they can be compared to each other without considerable refactoring. The biggest problem with scale indices is that new indices have been introduced and some of the source variables have been changed over the years. For these reasons only those scale indices were selected for this thesis where the original variables are the same.

It is stated in the technical report of 2012 (section Transforming the plausible values to PISA scales), that ”For PISA 2012, the reading, mathematics and science results are each reported on the scales that were established when the respective domain was a major domain. [–] In the case of mathematics they are directly comparable with the results reported in PISA 2003, 2006 and 2009”. (OECD 2014, p 253)

A total of 9 scale indices were selected. Five of them are related to students' attitudes towards mathematics, two are related to school and educational environment and two were related to student-teacher relations. All of the scale indices had a question that asked the student to assess the specific statements. Table 1 displays the statements and provided mathematical problems and has references to table 2 which shows the given questions and possible answers.

Scale index	Description	Statements	Q&A
ANXMAT (mathematics)	Anxiety towards mathematics	I often worry that it will be difficult for me in mathematics classes I get very tense when I have to do mathematics homework I get very nervous doing mathematics problems I feel helpless when doing a mathematics problem I worry that I will get poor <grades> in mathematics	1
ATSCHL (environment)	Attitude towards school (learning outcomes)	School has done little to prepare me for adult life when I leave school School has been a waste of time School has helped give me confidence to make decisions School has taught me things which could be useful in a job	1

DISCLIM(A) (environment)	Disciplinary climate	Students don’t listen to what the teacher says There is noise and disorder The teacher has to wait a long time for students to <quiet down> Students cannot work well Students don’t start working for a long time after the lesson begins	3
INTMAT (mathematics)	Interest towards mathematics	I enjoy reading about mathematics I look forward to my mathematics lessons I do mathematics because I enjoy it I am interested in the things I learn in mathematics	1
INSTMOT (mathematics)	Instrumental motivation for mathematics	Making an effort in mathematics is worth it because it will help me in the work that I want to do later on Learning mathematics is worthwhile for me because it will improve my career <prospects, chances> Mathematics is an important subject for me because I need it for what I want to study later on	1
MATHEFF (mathematics)	Mathematics self-efficacy	Using a <train timetable> to work out how long it would take to get from one place to another Calculating how much cheaper a TV would be after a 30% discount Calculating how many square meters of tiles you need to cover a floor Understanding graphics presented in newspapers Solving an equation like $3x+5=17$ Finding the actual distance between two places on a map with a 1:10 000 scale Solving an equation like $2(x+3)=(x+3)(x-3)$ Calculating the petrol consumption rate of a car	2
SCMAT (mathematics)	Mathematics self-concept	I am just not good at mathematics I get good <grades> in mathematics I learn mathematics quickly I have always believed that mathematics is one of my best subjects In my mathematics class, I understand even the most difficult work	1
STU(D)REL (relations)	Student-teacher relations	Students get along well with most teachers Most teachers are interested in students’ well-being Most of my teachers really listen to what I have to say If I need extra help, I will receive it from my teachers Most of my teachers treat me fairly	1

TEAHCSUP (relations)	Teacher support in mathematics classes	The teacher shows an interest in every student's learning The teacher gives extra help when students need it The teacher helps students with their learning The teacher continues teaching until the students understand The teacher gives students an opportunity to express opinions	3

Table 1: Mathematics scale indices

Reference number	Question	Possible answers
1	How much do you disagree or agree with the following statements about how you feel when studying mathematics?	Strongly agree Agree Disagree Strongly disagree
2	How confident do you feel about having to do the following calculations?	Very confident Confident Not very confident Not at all confident
3	How often do these things happen in your mathematics lessons?	Every lesson Most lessons Some lessons Never or hardly ever

Table 2: List of questions and possible answers related to scale indices

Once the decision was made about which variables should be used for clustering, the wanted columns were saved to a new data table. As mentioned in section 2.1, the 2012 data set was missing a significant amount of the values (see table 3) due to the rotation of the questions. Since k-means++ cannot handle rows with missing values (see section 3) this would be a massive problem, because in the 2012 data nearly every row was missing some values. The missing values needed to be replaced somehow; a method called imputation.

Scale index	Missing values 2003	%	Missing values 2012	%
ANXMAT	30	0.5	3141	35.6
ATSCHL	9	0.2	3163	35.8
DISCLIM(A)	55	0.9	3126	35.4
INSTMOT	102	1.8	3072	34.8
INTMAT	99	1.7	3071	34.8
MATHEFF	99	1.7	3071	34.8
SCMAT	30	0.5	3140	35.6
STU(D)REL	14	0.2	3140	35.6
TEACHSUP	53	0.9	3102	35.1

Table 3: Missing values per scale index

4.2 Imputation

There are many ways to handle missing values; (Luengo, García, and Herrera 2012) lists and tests 14 different methods found in different articles, including *Do Not Impute* where the missing values are left as is. These methods are tested on different data sets and evaluated. One of the methods is the *Concept Average Value for Numerical Attributes* (CMC) where each missing value is substituted with the average value of the class where the missing value belongs to. A more general version of this is to replace the missing value with the average of the whole data for the attribute in question. These two methods got the highest ranking with *1-NN* methods (connecting each dot to the closest cluster which is basically what k-means aims to do) and CMC was clearly the most efficient. (Luengo, García, and Herrera 2012)

Since the Finnish PISA data for 2012 was missing so many values, replacing them with the mean of the whole data would have created a considerable bias, and in the end CMC was used for this thesis. As stated above, CMC requires the missing values to belong to certain classes. The problem to be considered was how to have enough classes to reduce the created bias enough in imputation and how to keep the amount of classes small enough so that the computational aspect could be easily handled and the classes easily understood.

There are many possible ways to divide PISA data, for example according to gender or socio-economical background. In PISA data there is a readily created variable called STRATUM which divides the students into categories. The stratification was done for schools based on geographical location and then based on ”type of school, degree of urbanization, or minority composition” (OECD 2014, p 71). In 2003 data set there were 12 strata for Finnish students (ACER 2003) and in 2012 there were 17 strata for Finland (ACER 2012). Even though the number of strata differs between the data sets, the essence of the categorization is the same (to divide the students into groups based on certain characteristics) and thus the variable can be used for the imputation. For each stratum the average was calculated and each missing value in the scale indices was replaced with the average of the stratum the student belongs to.

As mentioned in chapter 2, the data represents the whole population only if the student weights are taken into consideration. Since the k-means++ which is built into Matlab does not handle weights, it could not be used. As stated in chapter 3, the script for modified k-means++ (see appendix A, listing 6.1) was provided by the thesis instructors.

4.3 Processing the data

The k-means++ script was tested with multiple values of k. For each k the function was called ten times and the results with the smallest error were saved for later comparison. The aim was to find the ideal amount of clusters for both data sets (hopefully the same number of clusters for both) so that the clustering error would be small and the distances between the clusters (more precisely the centroids of the clusters) would be big. The clustering error can be calculated as the sum of distances from each observed value to the closest centroid. Clustering error is zero when there are as many clusters as there are observed values, but when the amount of clusters increases, the clustering error gets smaller more slowly. When looking at the graphs of clustering error of each data set (see figure 3), it would seem that the clustering error starts to decrease noticeably slower at around three-cluster mark.

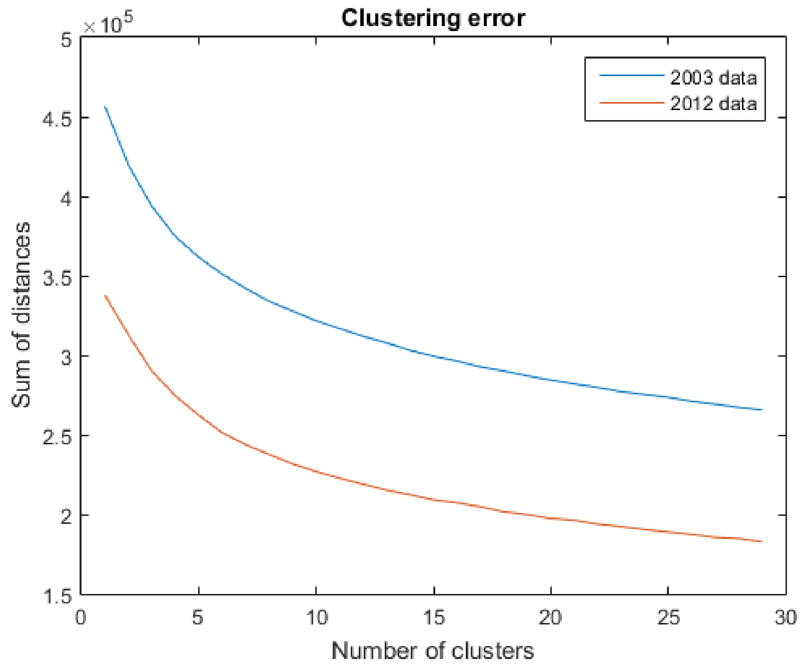

Figure 3. Original clustering errors

Since subjective interpretation of a single graph is not enough, additional information was needed. One way is to use indices that can be calculated to represent the quality of the solution for each k. In this thesis the *Ray-Turi index* (RTI) was used. The basic formula is: $validity = \frac{intra}{inter}$. Intra is the sum of distances between each point within a cluster (should be minimized), and inter is the distances between the clusters (should be maximized). The aim is to find the k for which the RTI is the smallest. (Ray and Turi 1999). Once the RTI had been calculated for all the k's used in the clustering, a graph was created for both data sets to represent the results (see figure 4).

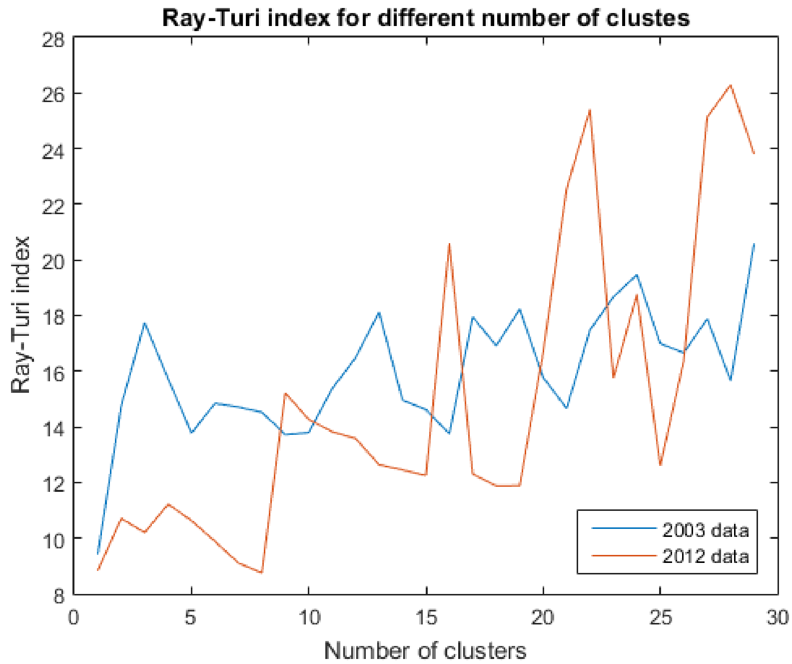

Figure 4. Ray-Turi index

The graph should have been a rather smooth rising curve, but the results that were received from these data sets seemed to differ from this a lot and the line representing the RTI jumped considerably in the graphs (see figure 4). This is why the values of each scale index were scaled to the range of [-1,1] to make sure that the data is as comparable as possible. It was also decided that there was no need to consider values of *k* higher than ten since interpreting the results gets challenging when the amount of clusters gets high. After the scaling had been done, the k-means++ was applied again and the RTI calculated. It was concluded that the nature of this data is such that the graphs for RTI will not be smooth (see figures 5, 6 and 7).

Based on the RTI graphs a good value for *k* in 2003 data would have been two or three, and for 2012 the best values would have been two or five. If only two clusters had been chosen for both data sets, the clusters would have been so big that clear differences would have been more challenging to find. Since the RTI for both 2003 and 2012 has a local minimum at five (see figure 5), it was decided to choose five clusters for both data sets to ease the comparison.

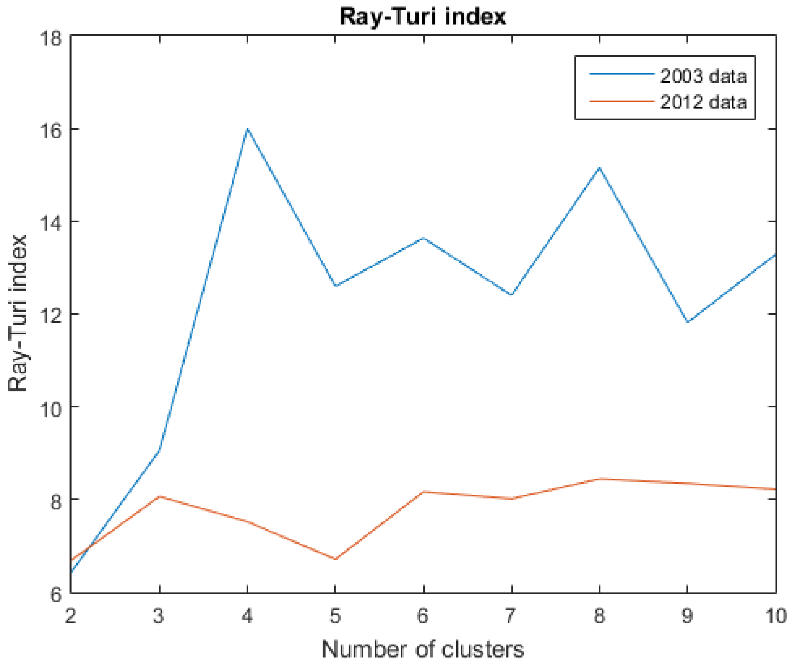

Figure 5. Ray-Turi index after scaling

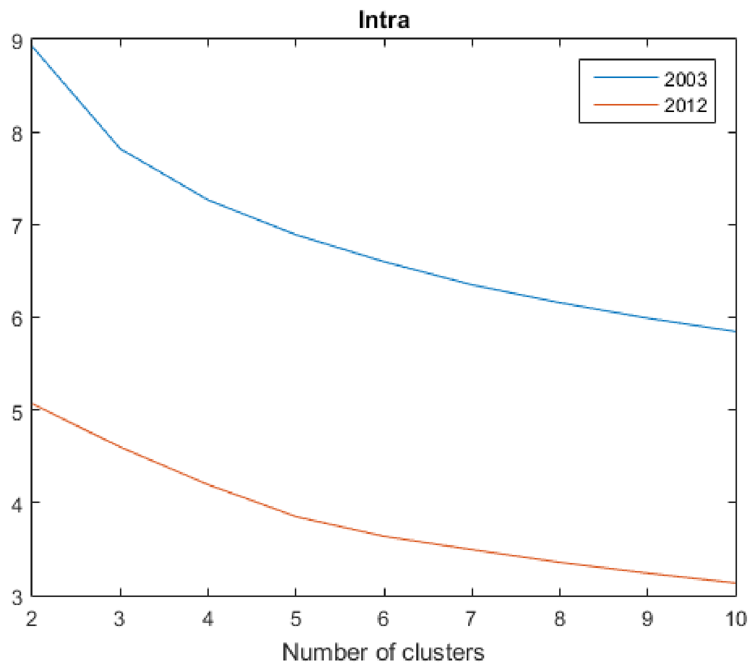

Figure 6. Intra for Ray-Turi index after scaling

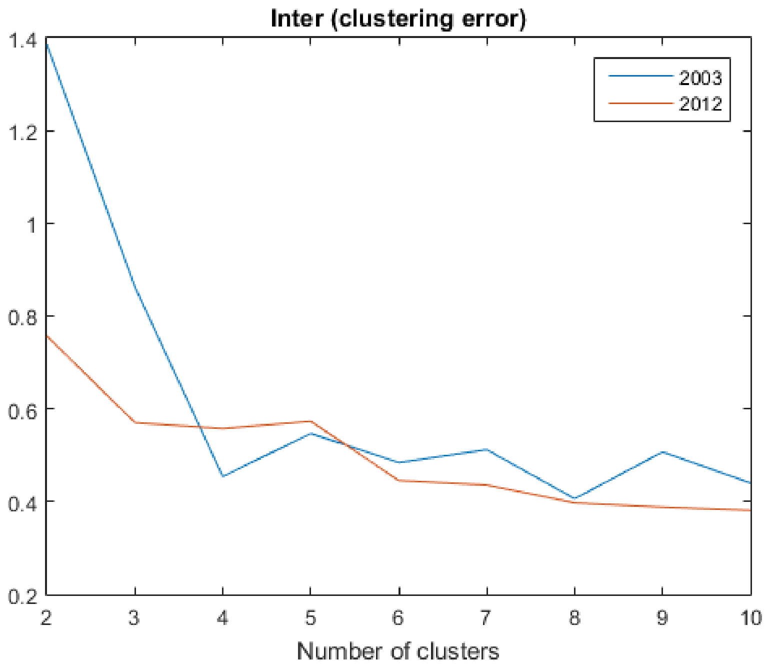

Figure 7. Inter for Ray-Turi index after scaling

5 Clusters

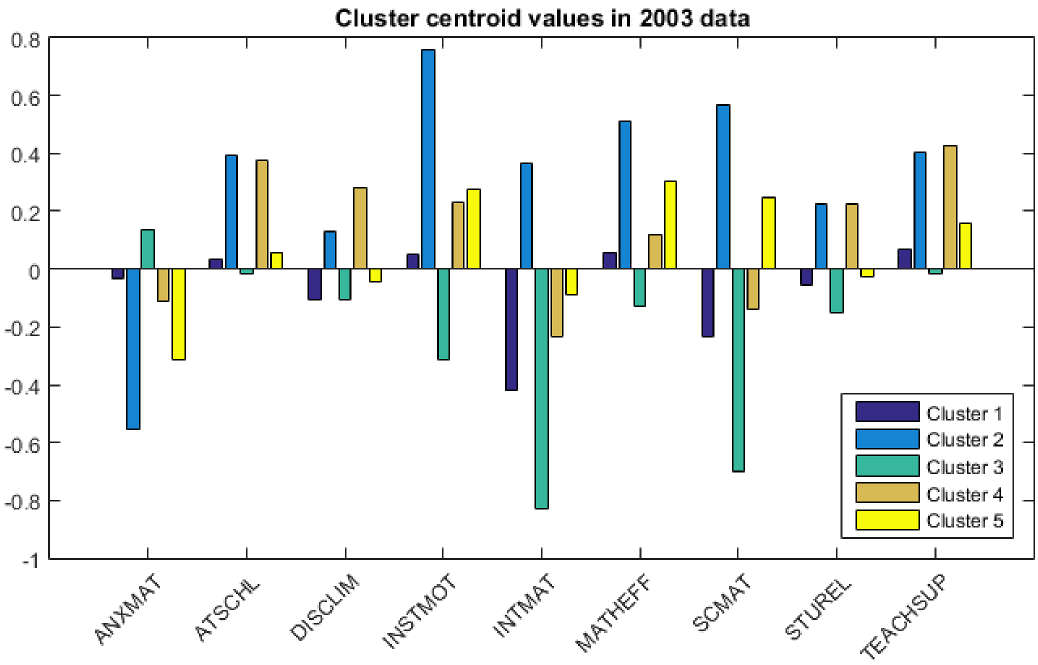

Figure 8. Cluster centroid values for 2003 data

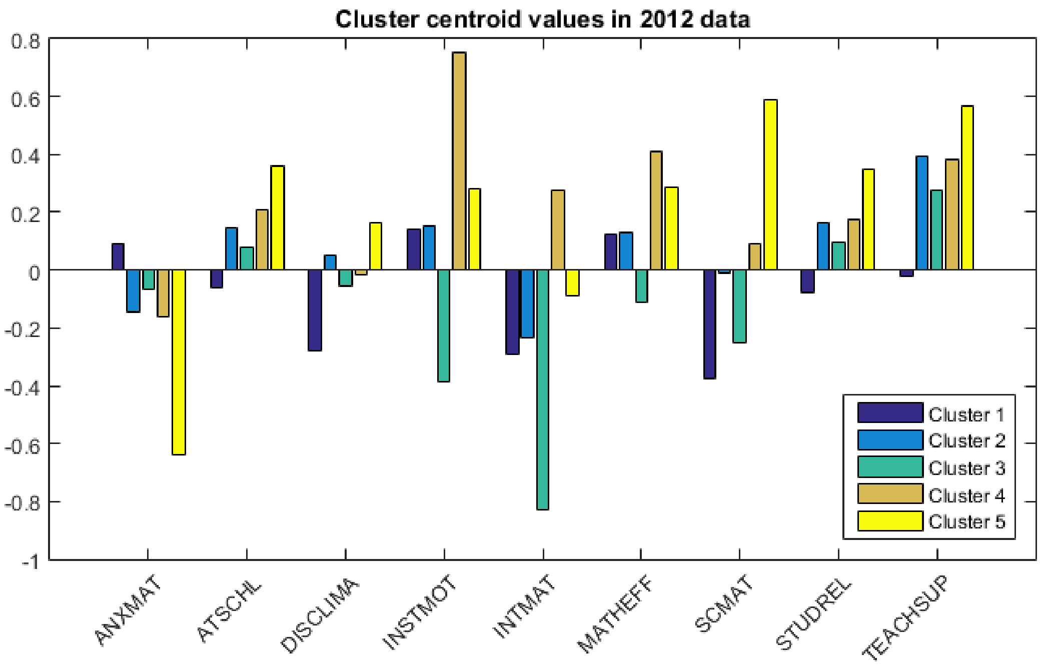

Figure 9. Cluster centroid values for 2012 data

This chapter describes the discovered clusters. Figures 8 and 9 show the values of the centroid of each cluster. The data has been scaled to the interval of $[-1,1]$ before clustering.

For both data sets a *proficiency level* has been given (for 2003 in table 5 and for 2012 in table 7. This proficiency level is based on *plausible values* that represent mathematical literacy. Table 4 shows the score points for each level for both data sets.

Level	2003	2012
6	Above 669	Above 669.3
5	607 - 669	607.0 to less than 669.3
4	545 - 607	544.7 to less than 607.0
3	482 - 545	482.4 to less than 544.7
2	420 - 482	420.1 to less than 482.4
1	358 - 420	357.8 to less than 420.1
Below 1	Not defined	Below 357.8

Table 4: The scores for each proficiency level

5.1 2003 data

Cluster	N	♀	♂	PV1M	PV2M	PV3M	PV4M	PV5M	Mean	Level
1	1546	803	743	516.3	515.9	516.3	516.9	516.2	516.3	3
2	862	303	559	607.8	608.3	608.1	609.5	610.2	608.8	5
3	872	585	287	486.8	486.4	487.4	488.4	485.6	486.9	3
4	1106	681	425	528.7	528.2	527.5	526.6	528.7	527.9	3
5	1410	557	853	577.6	577.8	577.3	579.6	579.4	578.3	4

Table 5: Basic information on 2003 clusters

Five plausible values (see section 2.1) related to mathematics literacy have been calculated. In this thesis the five main PVs for mathematics have been used. Table 5 shows the amount of students in each cluster, gender distribution and weighted means of the five plausible values.

It also shows the average of the plausible values for each cluster and shows which proficiency level the cluster belongs to (see also table 4). Table 6 displays the weighted means of each scale index for all students and for each cluster separately.

Scale index	All students	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
ANXMAT	-0.32	0.01	-1.35	0.45	-0.20	-0.72
ATSCHL	0.11	-0.23	0.80	-0.36	0.76	-0.16
DISCLIM	-0.15	-0.45	0.20	-0.43	0.55	-0.28
INSTMOT	0.06	-0.23	1.24	-0.99	0.15	0.24
INTMAT	-0.24	-0.56	1.08	-1.45	-0.16	0.16
MATHEFF	-0.15	-0.50	0.97	-1.09	-0.28	0.32
SCMAT	0.01	-0.38	1.39	-1.44	-0.17	0.68
STUREL	-0.03	-0.28	0.55	-0.56	0.58	-0.19
TEACHSUP	0.08	-0.24	0.61	-0.45	0.66	-0.01

Table 6: Weighted means in 2003 clusters

In the 2003 data there are 5796 Finnish students. After the clustering was done, cluster 1 had 1546 students which is about 27% of the students. There were slightly more female than male students. Cluster 1 was close to the average in regards to most of the scale indices used in clustering. Attitude towards school was the second worst of all the clusters. The students in cluster 1 had the second lowest experience on teacher support and student-teacher relationship. Along with cluster 3 they had the worst experience in the disciplinary climate. Anxiety towards mathematics was quite high. Instrumental motivation and students' own motivation towards mathematics was the second lowest as well as their self-concept towards mathematics. Their confidence in solving the given problems was also the second lowest. The students in cluster 1 got the second lowest score in mathematical literacy (see "Mean" in table 5).

Cluster 2 had 862 students which is 15% of students. 303 (35%) were female and 559 (65%) male. Cluster 2 stood out in a positive way in most scale indices. Cluster 2 had the best attitude towards school, they experienced positive relations with teachers and got a high level of teacher support. Their experience of the disciplinary climate was clearly positive. The students in cluster 2 had the lowest anxiety towards mathematics and distinctly the highest motivation (both personal and instrumental) towards mathematics. They were most cofident about being able to solve the mathematical problems and had the highest self-concept in mathematics. Cluster 2 was clearly the most mathematically-oriented cluster and the students also seemed to like the school environment and got along well with the mathematics teachers. They got the highest score in mathematical literacy.

Cluster 3 had 872 students which is 15% of students. 585 (67%) were female and 287 (33%) male. Cluster 3 stood out in a negative way in most scale indices. The students in cluster 3 had the most negative attitude towards school, the most negative student-teacher relations and their experience on teacher support was negative. They also experienced the disciplinary climate as negative as the students in cluster 1. Cluster 3 clearly had the highest anxiety towards mathematics, a significantly low trust in their mathematics skills and the lowest instrumental and personal motivation towards mathematics. They also had the lowest confidence in their ability in solving the given mathematical problems. The students in cluster 3 got the lowest score in mathematical literacy.

Cluster 4 had 1106 students which is 19% of students. 681 (62%) were female and 425 (38%) male. The students in cluster 4 had a very positive attitude towards school and good relations with the teachers. For them the disciplinary climate was the most positive and they were most confident on getting help from the teacher. Anxiety towards mathematics was lower than in clusters 1 and 3 but higher than in clusters 2 and 5. Cluster 4 was also in between of the other four clusters in other math related scale indices. The students in this cluster seemed to think that mathematic skills would be useful in the future (instrumental motivation), but did not have a very high personal interest in mathematics. Cluster 4 did not stand out in any of the mathematics-related scale indices nor mathematical literacy even though they seemed to like school and the teachers.

Cluster 5 had 1410 students which is 24% of students. 557 (40%) were female and 853 (60%) male. In the scale indices related to school and teachers, cluster 5 was between the other clusters (two having more positive attitudes and two more negative attitudes). The students in cluster 5 had a low level of anxiety towards mathematics, had the second highest instrumental motivation, a good self-concept in mathematics and a high confidence in being able to solve the given tasks (but not as high as in cluster 2). Their personal interest in mathematics was not particularly high, but it was the second highest of all clusters. Cluster 5 seemed to be mathematically oriented even though their attitude towards school and teachers was not as high as in clusters 2 and 4. The score for mathematical literacy was the second highest.

5.2 2012 data

Cluster	N	♀	♂	PV1M	PV2M	PV3M	PV4M	PV5M	Mean	Level
1	1272	720	552	475.4	475.6	475.1	473.2	475.8	475.0	2
2	4191	2056	2135	495.6	496.7	496.6	495.9	496.2	496.2	3
3	966	564	402	466.7	464.9	466.4	466.0	465.4	465.9	2
4	1302	580	722	543.4	541.7	544.0	543.5	541.5	542.8	3
5	1098	450	648	579.6	578.9	579.2	579.6	580.6	579.6	4

Table 7: Basic information on 2012 clusters

Scale index	All students	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
ANXMAT	-0.33	0.39	-0.19	0.17	-0.23	-1.43
ATSCHL	0.06	-0.47	0.06	-0.25	0.52	0.67
DISCLIMA	-0.33	-0.93	-0.10	-0.55	-0.28	0.11
INSTMOT	-0.01	-0.24	-0.09	-1.08	1.12	0.44
INTMAT	-0.22	-0.57	-0.19	-1.41	0.88	0.49
MATHEFF	-0.27	-0.60	-0.37	-1.08	0.42	0.65
SCMAT	0.03	-0.85	-0.00	-1.06	0.64	1.30
STUDREL	-0.09	-0.67	0.04	-0.27	0.22	0.51
TEACHSUP	0.16	-0.66	0.39	-0.08	0.45	0.76

Table 8: Weighted means in 2012 clusters

In the 2012 data there are 8829 Finnish students. Table 7 displays the basic data on the 2012 clusters, including the amount of students in each cluster and the plausible values related to mathematics. Table 8 presents the weighted means of the scale indices for all Finnish students and for each of the clusters. After the clustering was done, cluster 1 had 1272 students which is about 14% of the students. There were 720 (57%) female and 552 (43%) male students. The students in cluster 1 had the most negative attitude towards school, experienced the disciplinary climate clearly in a more negative way than others, had the lowest confidence on getting help from the teacher and had the most negative student-teacher relations. Students in cluster 1 had the highest anxiety towards mathematics and very low confidence in their own abilities in mathematics. Instrumental motivation towards mathematics was close to average, but personal interest was the second lowest. Cluster 1 did not stand out in any way in assessing whether or not they would be able to solve the given tasks. The students in cluster 1 got the second lowest score in mathematical literacy (see ”Mean” in table 7)

Cluster 2 was clearly the largest cluster. There were 4191 students which is 48% of all of the students. 2056 (49%) were female and 2135 (51%) male. It was average on most respects. Experience on disciplinary climate was the second best and experience on teacher support and student-teacher relations was positive. In cluster 2, anxiety towards mathematics was not very high (not the lowest either), instrumental motivation was in between other clusters as well as mathematics self-efficacy and self-concept in mathematics. Personal interest in mathematics was rather low but still between the other clusters.

Cluster 3 had 966 students which is 11% of students. 564 (58%) were female and 402 (42%) male. Students in cluster 3 had the second lowest values in all of the scale indices related to school environment and teachers and the lowest values in almost all of the mathematics-related scale indices. Only in mathematics self-concept cluster 1 had a more negative estimate. For cluster 3 the weighted mean (see table 8) in all mathematics-related scale indices were the lowest except for anxiety towards mathematics where it was the second highest. Cluster 3 was the least mathematics-oriented cluster, but the students' experience on school and teachers was not as negative as in cluster 1. Cluster 3 had the lowest score in mathematical literacy.

Cluster 4 had 1302 students which is 15% of students. 580 (45%) were female and 722 (55%)

male. Their attitude towards school was positive (the second highest), they had good student-teacher relations and were confident on getting help from the teacher when needed. Their experience on the disciplinary climate was average (between the other clusters). Cluster 4 had reasonably low anxiety towards mathematics. They had distinctively higher instrumental motivation and personal interest towards mathematics than the other clusters. Students in cluster 4 were the most confident that they would be able to solve the given mathematical problems, but had only the second highest mathematics self-concept where the difference to cluster 5 was significant. The score in mathematical literacy was the second highest.

Cluster 5 had 1098 students which is 12% of students. 450 (41%) were female and 648 (59%) male. Cluster 5 had the most positive attitude towards school and the most positive experience on the disciplinary climate. They also had the most positive student-teacher relations and were most confident on teacher support in mathematics classes. Cluster 5 had clearly the lowest anxiety towards mathematics. They had the second highest instrumental motivation and personal interest towards mathematics and were the second most confident on being able to solve the presented mathematical tasks. They had clearly the most positive mathematics self-concept. The students in cluster 5 had the highest score in mathematical literacy.

5.3 Comparison

There are certain similarities between the clusters of the two data sets. There are clusters that stand out in the mathematical tendencies and scores and clusters that stand out by having the lowest interest and lowest scores. However, there are some minor changes in what the clusters represent.

In 2003 data none of the clusters were level 1 or 2 in mathematics proficiency (although cluster 3 was close to level 2) nor did they reach level 6. In 2012 data two of the clusters were level 2 in proficiency and none of the clusters reached level 5 or 6. This complies with the findings that Finland's ranking in mathematics (in PISA studies) has dropped, and even in national level the mathematics-related results have dropped significantly (*"PISA 12 Still among the best in the OECD but performance is declining"* n.d.).

In 2003, cluster 2 got the highest score in mathematical literacy, and in 2012 cluster 5 got the highest score. In both clusters the school- and teacher-related indices got high scores. In 2003 (cluster 2) the indices related to mathematics had the highest scores, but in 2012 only anxiety towards mathematics was the lowest and the mathematics self-concept was considerably higher than in other clusters. In personal and instrumental motivation as well as mathematics self-efficacy cluster 5 (2012) was only the second highest.

It seems that in 2012, the students who have the highest instrumental motivation, the highest interest in mathematics and the highest mathematical self-efficacy (cluster 4 in 2012) no longer receive the highest score in mathematics. Instead it looks like what mattered more in 2012 was low anxiety towards mathematics, good student-teacher relations and confidence on getting help when needed. These students also had a stronger belief that they could solve the given mathematical tasks. Perhaps the reason why cluster 5 did not reach level 5 in mathematical proficiency is that they had a much lower motivation towards mathematics than the students in cluster 4. They did not seem to think that mathematics is as interesting or needed as much as the students in cluster 4.

When comparing the clusters that received the lowest score in plausible values, an interesting difference can be found; In 2003 cluster 3 was the only one that received a score below 500 and the difference to the second lowest cluster was 30 points. In 2012 there were two clusters (1 and 3) that received scores that were below 500 and the difference between them is only 10 points. They both got lower scores than cluster 3 in 2003. In addition cluster 2 which represents the so called middle group in 2012 (when looking at the plausible values), got only 10 points more than cluster 3 in 2003. Cluster 2 in 2012 had the majority of the students in that dataset, and so it is worrying that the mean of the plausible values for this cluster is so low.

In 2003 cluster 3 got negative values for all of the scale indices except for anxiety towards mathematics where it was the only group that had a positive value (meaning there was high level of anxiety towards mathematics). In 2012 all of the clusters got positive values in at least a couple of the other indices, so based on the scale indices there is no longer a cluster that would have very ”bad” attitudes towards both mathematics and school in general.

Cluster 3 in 2012 stood out in a negative way in mathematics-related scale indices which is probably the reason why they got the lowest score in plausible values of mathematical literacy. However, they had positive values in school-related scale indices. It might give more information about this particular cluster to find out in future research what type of results this cluster got in subjects other than mathematics. Cluster 1 in 2012 got mostly negative values, but in anxiety towards mathematics, instrumental motivation and mathematics self-efficacy they had positive values. This means that even though they did not seem to like school or the teachers very much and they had the highest anxiety towards mathematics, they believed that mathematics is needed in life and they were reasonably confident on being able to solve the given mathematical tasks. For clusters 1 and 3 in 2012 the disadvantages seem to be opposites: in cluster 1 the students seemed to have a very negative attitude towards school but less negativity towards mathematics, and in cluster 3 the students had a very negative attitude towards mathematics but less negative attitude towards school.

Cluster 4 in 2003 and cluster 2 in 2012 represent the middle group when looking at the plausible values. They are also in the middle ground in the scale indices. The biggest differences between these two clusters are that when moving from 2003 to 2012, the scores in attitude towards school, disciplinary climate and instrumental motivation have dropped. There is also a small drop in student-teacher relations, but self-concept in mathematics has gone up slightly.

		2012				
		1	2	3	4	5
2003	1	0.34	0.55	0.67	1.19	1.36
	2	1.71	1.20	2.05	0.68	0.73
	3	0.84	1.23	0.64	1.92	2.03
	4	0.96	0.37	1.02	0.89	0.95
	5	0.87	0.52	1.22	0.73	0.82

Table 9: Distances between 2003 and 2012 centroids

The distances between the centroids of each cluster was calculated to back up the initial analysis that was based on the characteristics of the clusters. According to table 9, cluster 4 in 2012 and cluster 2 in 2012 are very close which means that they can be considered as similar groups and can be compared as such. The smallest distance is between cluster 1 in 2003 and cluster 1 in 2012.

Clusters in 2003 from highest to lowest score	Corresponding clusters in 2012 based on score	Corresponding clusters in 2012 based on distance (scale indices)
2 (highest)	5	4
5	4	5
4	2	2
1	1	1
3 (lowest)	3	3

Table 10: Connections between the datasets

Table 10 shows the 2003 clusters in order from highest to lowest based on the plausible values used to evaluate mathematical literacy (see table 5). In the second column the clusters of 2012 are in a similar order from highest to lowest score in mathematical literacy. The third column shows the clusters of 2012 so that the cluster on the first row (cluster 4) resembles the cluster of 2003 that is on the same row (cluster 2) based on the students' attitudes towards school and mathematics. One should pay attention to the differences between the second and third column.

When looking at table 10 it can be seen that the clusters receiving the lowest scores have similar characteristics. The clusters having the second lowest score and medium score are also similar in 2003 and 2012. In the higher scores there have been some changes. When looking at similarities in students' tendencies, cluster 4 in 2012 has the most resemblance to cluster 2 in 2003 which received the highest score (see table 10). However, in 2012 cluster 4 got only the second highest score. Cluster 5 in 2012 received the highest score, but the score was considerably lower than the highest score in 2003 and actually closer to the second highest score in 2003. Cluster 5 in 2012 actually seems to be closer to the second highest cluster (cluster 5) in 2003 concerning the student's tendencies.

It is clear that the students receiving lower scores in mathematical literacy had similar profiles in 2003 and 2012. In addition, there was a cluster in 2003 that had students who were interested in mathematics and liked school. Such cluster can no longer be found in 2012. Instead there is a cluster (cluster 5) where students had the most positive attitude towards school and teachers, but did not have notably high motivation towards mathematics. Interestingly they had the lowest anxiety towards mathematics and received the highest score in mathematical literacy. There is another cluster (cluster 4) in 2012 where the students had considerably high motivation towards mathematics and had a high belief they could solve the given tasks, but their self-efficacy in mathematics was low and anxiety towards mathematics was high when compared to cluster 5. Cluster 4 got the second highest score.

It would seem that how students rate the school environment has a higher impact on the performance (at least in mathematics) than interest in the subject itself. In 2003 attitudes towards the school environment and mathematics were closely related so that if one was low, the other one was low as well. In 2012 this relation had gotten weaker in both the highest and the lowest scoring clusters. The students who had the most negative attitude towards school no longer had the lowest interest/motivation towards mathematics. In the same way the students who had the most positive attitude towards school no longer had the highest interest/motivation towards mathematics.

6 Conclusion

For a long time people have been interested in assessing student performance and learning, and comparing the results internationally. In 2000 OECD countries started a program for international student assessment (PISA) which has since become a well-known and well-structured way for student assessment. Finland has received top results in the PISA studies from 2000 to 2009. Even though the results were still good in 2012, there had been a significant drop in Finland's results. Since the 2012 PISA study focused on mathematics, its results were compared to 2003, the previous PISA study with the same focus. Only those scale indices (readily scaled variables consisting of more than one of the original variables) were used which had remained the same.

The used method was k-means++, which is a data mining algorithm aiming to divide the data set into k clusters. Student weights were implemented into the algorithm so that the results would represent the whole population of 15-yeal-old students in Finland. After running the algorithm multiple times with different values for k, an additional index, called Ray-Turi index (RTI), was calculated. Based on the clustering error and RTI, it was decided that five clusters for both data sets would be the best.

The found clusters were analyzed based on the average of the scaled indices and additional variables, such as gender. Most of the clusters in 2012 data correspond with those that were found in (Saarela and Kärkkäinen 2014). When comparing the clusters of 2003 and 2012, it is clear that the clusters which got the lowest results resemble each other greatly. A notable difference is that the lowest performing cluster in 2003 had negative attitudes towards school, teacher and mathematics, but in 2012 the students who had the most negative attitudes towards mathematics were no longer the most negative towards school and teachers and did not have the highest anxiety towards mathematics. Instead the cluster which had the highest anxiety in 2012 had the most negative attitudes towards teachers and school, but seemed to be more confident about their skills and the importance of mathematics.

It is noteworthy that in 2012 there was one notably large cluster which did not stand out in the scale indices or the plausible values. In both data sets the found clusters had between

800 and 1400 students, but the biggest cluster in 2012 had over 4000 students, which was unexpected. It is not clear why this happened, and it would be a topic for further studies to find out whether this was due to selecting wrong amount of clusters or if this has something to do with the variables that were used in the study.

In the clusters that got the highest results there are clear differences; in 2003 the cluster which received the best results had students who had a good attitude towards school and had personal interest and motivation towards mathematics. In 2012 there were two top clusters. One had students who had a good attitude towards school and got good results in mathematics, but did not have a lot of personal interest towards mathematics. The other cluster had students who were interested in and motivated towards mathematics, but did not have such a positive attitude towards school. All of the clusters in 2012 received lower results than the corresponding clusters in 2003.

Based on the results it would seem that in 2012 there have been two types of well-performing students: those who like school in general and those who are more interested in the subject itself. This may be an indication of a change in attitudes towards school, or it may be an indication that something has happened that has changed the school environment.

For further studies, more variables should be included in the clustering and analysis. It would also be good to study the performance and attitudes in other domains (reading and science) to gain more information about whether these results are specific to mathematics or if there has been a bigger change.

Bibliography

"About PISA". n.d. https://www.oecd.org/pisa/aboutpisa/. Accessed: 2016-11-01.

ACER. 2003. *Codebook Student Questionnaire.* http://pisa2003.acer.edu.au/downloads/StQ_CodeBook_2003.pdf.

———. 2012. *Codebook for PISA 2012 Main Study Student Questionnaire: MAIN DATABASE.* http://pisa2012.acer.edu.au/downloads/M_stu_codebook.pdf.

Arthur, David, and Sergei Vassilvitskii. 2007. "k-means++: The advantages of careful seeding". In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,* 1027–1035. Society for Industrial and Applied Mathematics.

"Basic Education Act 628/1998". n.d. http://www.finlex.fi/en/laki/kaannokset/1998/en19980628.pdf.

De Corte, Erik, Rianne Janssen, and Lieven Verschaffel. 2009. "Large-Scale Assessment as a Tool for Monitoring Learning and Teaching: The Case of Flanders, Belgium".

Economic Co-operation, Organisation for, and Development. 2003. *Literacy skills for the world of tomorrow: Further results from PISA 2000.* OECD Publishing.

Education, Ministry of, and Culture. n.d.(a). "OKM - Background for Finnish PISA success". Accessed: 2016-09-05. http://www.minedu.fi/pisa/taustaa.html?lang=en.

———. n.d.(b). "OKM - The Results of PISA 2000". Accessed: 2016-10-25. http://www.minedu.fi/pisa/2000.html?lang=en.

———. n.d.(c). "OKM - The Results of PISA 2003". Accessed: 2016-10-25. http://www.minedu.fi/pisa/2003.html?lang=en.

———. n.d.(d). "OKM - The Results of PISA 2009". Accessed: 2016-10-25. http://www.minedu.fi/pisa/2009.html?lang=en.

Education, Ministry of, and Culture. n.d.(e). “OKM - The Results of PISA 2012”. Accessed: 2016-10-25. http://www.minedu.fi/pisa/2012.html?lang=en.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. “From data mining to knowledge discovery in databases”. *AI magazine* 17 (3): 37.

Hoepfl, Marie. 2000. “Large-scale authentic assessment”. *Using authentic assessment in vocational education:* 49–67.

Kupiainen, Sirkku, Jarkko Hautamäki, and Tommi Karjalainen. 2009. *The Finnish education system and PISA.* Volume 46. Ministry of Education Helsinki.

Luengo, Julián, Salvador García, and Francisco Herrera. 2012. “On the choice of the best imputation methods for missing values considering three groups of classification methods”. *Knowledge and information systems* 32 (1): 77–108.

OECD. 2005. *PISA 2003 Technical Report.* https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35188570.pdf.

———. 2014. *PISA 2012 Technical Report.* https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

”PISA 12 Still among the best in the OECD but performance is declining”. n.d. http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2013/liitteet/PISA12_leaflet.pdf?lang=fi. Accessed: 2016-09-05.

PISA, OECD. 2012. *Results in Focus: What 15-year-olds know and what they can do with what they know.*

Ray, Siddheswar, and Rose H Turi. 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,* 137–143. Calcutta, India.

Saarela, Mirka, and Tommi Kärkkäinen. 2014. “Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices”. In *Proceedings of the 7th International Conference on Educational Data Mining,* 60–68.

Saarela, Mirka, and Tommi Kärkkäinen. 2015. “Weighted Clustering of Sparse Educational Data”. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning -ESANN 2015,* 337–342.

Shavelson, Richard J, Carol Geary Schneider, and Lee S Shulman. 2007. *A brief history of student learning assessment: How we got where we are and a proposal for where to go next.* Association of American Colleges / Universities.

Välijärvi, Jouni, Pekka Kupari, Pirjo Linnakylä, Pasi Reinikainen, Sari Sulkunen, Jukka Törnroos, and Inga Arffman. 2007. *The Finnish success in Pisa-and some reasons behind it: Pisa 2003. 2.* Jyväskylän yliopisto, Koulutuksen tutkimuslaitos.

Zaki, Mohammed J, and Wagner Meira Jr. 2014. *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.

Appendices

A Codes

Listing 6.1. K-means++ algorithm that takes weights into consideration

```
function [L,C] = kmeanspp(X,k,w)
%KMEANS Cluster multivariate data using the k-means++ algorithm.
%   [L,C] = kmeans(X,k) produces a 1-by-size(X,2) vector L with one class
%   label per column in X and a size(X,1)-by-k matrix C containing the
%   centers corresponding to each class.

%   Version: 2013-02-08
%   Authors: Laurent Sorber (Laurent.Sorber@cs.kuleuven.be)
%
%   References:
%   [1] J. B. MacQueen, "Some Methods for Classification and Analysis of
%       MultiVariate Observations", in Proc. of the fifth Berkeley
%       Symposium on Mathematical Statistics and Probability, L. M. L. Cam
%       and J. Neyman, eds., vol. 1, UC Press, 1967, pp. 281-297.
%   [2] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of
%       Careful Seeding", Technical Report 2006-13, Stanford InfoLab, 2006.
%
% NOTE: The use of (N,1)-vector of weights added by Tommi K. Mar 11, 2016
%

L = [];
L1 = 0;

while length(unique(L)) ~= k

    % The k-means++ initialization.
    C = X(:,1+round(rand*(size(X,2)-1)));
    L = ones(1,size(X,2));
    for i = 2:k
        D = X-C(:,L);
        D = cumsum(sqrt(dot(D,D,1)));
```

```
        if D(end) == 0, C(:,i:k) = X(:,ones(1,k-i+1)); return; end
        C(:,i) = X(:,find(rand < D/D(end),1));
        [~,L] = max(bsxfun(@minus,2*real(C'*X),dot(C,C,1).'));
    end

    % The k-means algorithm.
    while any(L ~= L1)
        L1 = L;
        for i = 1:k
            l = L==i;
            %C(:,i) = sum(X(:,l),2)/sum(l); %Original formula
            C(:,i) = sum(bsxfun(@times,X(:,l),w(l)'),2)/sum(w(l));
        end
        [~,L] = max(bsxfun(@minus,2*real(C'*X),dot(C,C,1).'),[],1);
    end

end
```