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:

$$\mathcal{J}_w = \sum_{j=1}^{K} \sum_{i=1}^{n_j} w_i \|\text{Diag}\{p_i\}(x_i - c_j)\|_2. \tag{2}$$

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 $v = \big[8\ 8\ 8\ 8\ 9\ 9\ 9\ 9\ 9\ 9\big]$, 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.

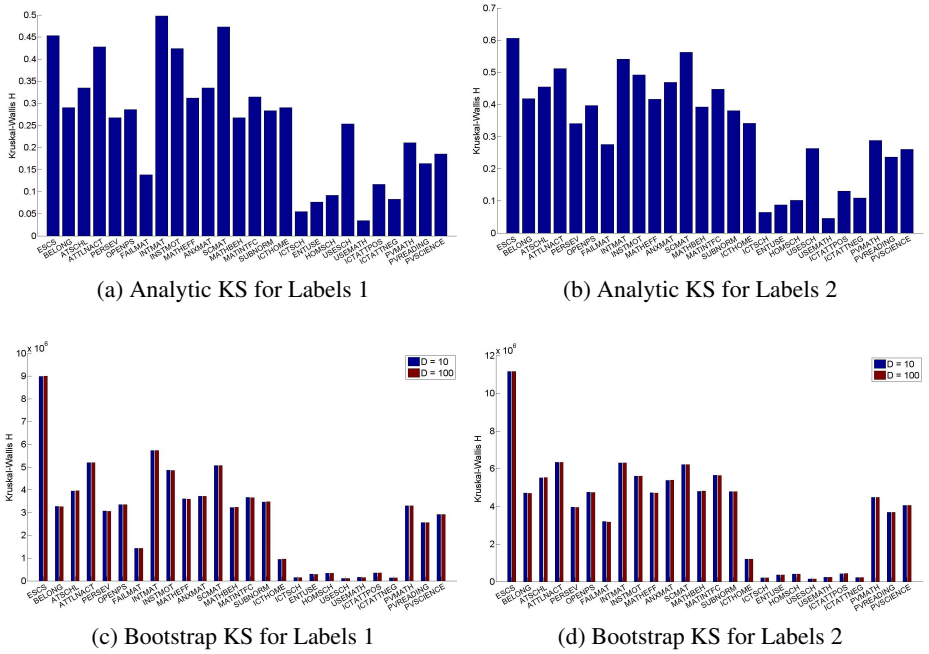
The five individual PVs for mathematics, reading, and science, as given in Table 1, were first treated as independent variables, such that five H values were computed for them. The final value of the test statistic was then taken as the mean of these according to the recommended way of analysis in [17].

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 $\times$ speedup compared to sequential code for Labels 1 and 35 $\times$ speedup for Labels 2. Correspondingly, parallel Bootstrap KW with $D = 10$ gives 28 $\times$ speedup for Labels 1 and 33 $\times$ 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)