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

Mirka Saarela
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
mirka.saarela@gmail.com

Tommi Kärkkäinen
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
tommi.karkkainen@jyu.fi

ABSTRACT

Certain stereotypes can be associated with people from different countries. For example, the Italians are expected to be emotional, the Germans functional, and the Chinese hard-working. In this study, we cluster all 15-year-old students representing the 68 different nations and territories that participated in the latest Programme for International Student Assessment (PISA 2012). The hypothesis is that the students will start to form their own country groups when clustered according to the scale indices that summarize many of the students' characteristics. In order to meet PISA data analysis requirements, we use a novel combination of our previously published algorithmic components to realize a weighted sparse data clustering approach. This enables us to work with around half a million observations with large number of missing values, which represent the population of more than 24 million students globally. Three internal cluster indices suitable for sparse data are used to determine the number of clusters and the whole procedure is repeated recursively to end up with a set of clusters on three different refinement levels. The results show that our final clusters can indeed be explained by the actual student performance but only to a marginal degree by the country.

Keywords

Weighted Clustering, PISA, Sparse Cluster Indices, Country Stereotype

1. INTRODUCTION

Certain stereotypes seem to be associated with people from different countries. The French and Italians, for example, are expected to be emotional, while Germany has mainly a functional country stereotype [4], and the Chinese are commonly perceived as hard-working [3]. According to the *Hofstede Model* [6], national cultures can be characterized along six dimensions: power distance, individualism, masculinity, uncertainty avoidance, pragmatism, and indulgence. The hypothesis in this study is that also the population of 15-year-old students worldwide will start to form their own national groups, i.e., show similar characteristics to their country peers, when clustered according to their attributes and attitudes towards education.

PISA (Programme for International Student Assessment) is a worldwide triannual survey conducted by the Organisation for Economic Co-operation and Development (OECD), assessing the proficiency of 15-year-old students from different countries and economies in three domains: reading, mathematics, and science. Besides evaluating student performances, PISA is also one of the largest public databases[1] of students' demographic and contextual data, such as their attitudes and behaviours towards various aspects of education.

In order to test our hypothesis, we utilize the 15 PISA scale indices (explicitly detailed in [14]), a set of derived variables that readily summarize the background of the students including their characteristics and attitudes. In particular, the *escs* index measures the students' economic, social and cultural status and is known to account for most variance in performance [9]. Additionally, 5 scale indices (*belong*, *atschl*, *attlnact*, *persev*, *openps*) are generally associated with performance on a student-level, while 9 further ones (*failmat*, *intmat*, *instmot*, *matheff*, *anxmat*, *scmat*, *mathbeh*, *matintfc*, *subnorm*) are directly related to attitudes towards mathematics, the main assessment area in the most recent survey (PISA 2012). However, since the assessment material exceeds the time that is allocated for the test, each student is administered solely a fraction of the whole set of cognitive items and only one of the three background questionnaires. Because of this rotated design, 33.24% of the PISA scale indices values are missing.

Moreover, PISA data are an important example of large data sets that include weights. Only some students from each country are sampled for the study, but multiplied with their respective weights they should represent the whole 15-year-old student population. The sample data of the latest PISA assessment, i.e., the data we are working with, consists of 485490 students which, taking the weights into account, represent more than 24 million 15-year-old students in the 68 different territories that participated in PISA 2012.

[1]See `http://www.oecd.org/pisa/pisaproducts/`.

The content of this paper is as follows. First, we describe the clustering algorithm that allows us to work with the large, sparse and weighted data (Sec. 2). Second, we present the clustering results (Sec. 3) and their relevance to our hypothesis, i.e., how the clusters on the different levels can be characterized and to what extent they form their own country groups. Finally, in Sec. 4, we conclude our study and discuss directions for further research.

2. THE CLUSTERING APPROACH

Sparsity of PISA data must be taken into account when selecting or developing a data mining technique. With missing values one faces difficulties in justifying assumptions on data or error normality [14, 15], which underlie the classical second-order statistics. Hence, the data mining techniques here are based on the so-called nonparametric, robust statistics [5]. A robust, weighted clustering approach suitable for data sets with a large portion of missing values, non-normal error distribution, and given alignment between a sample and the population through weights, was introduced and tested in [16]. Here, we apply a similar method with slight modifications, along the lines of [7] for sampled initialization and [17] for hierarchical application. All computations were implemented and realized in Matlab R2014a.

2.1 Basic method

Denote by N the number of observations and by n the dimension of an observation of the data matrix $\mathbf{X}$; and let $\{w_i\}, i = 1, \ldots, N$ be the positive sample-population-alignment weights. Further, let $\{\mathbf{p}_i\}, i = 1, \ldots, N$, be the projection vectors that define the pattern of the available values [10, 1, 14, 15]. The weighted spatial median $\mathbf{s}$ with the so-called available data strategy can be obtained as the solution of the projected Weber problem

$$\min_{\mathbf{v}\in\mathbf{R}^n} \mathcal{J}(\mathbf{v}), \quad \mathcal{J}(\mathbf{v}) = \sum_{i=1}^{N} w_i \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{v})\|, \qquad (1)$$

where $\mathrm{Diag}\{\mathbf{p}_i\}$ denotes the diagonal matrix corresponding to the given vector $\mathbf{p}_i$. As described in [8], this optimization problem is nonsmooth, i.e., it is not classically differentiable. However, an accurate approximation for the solution of the nonsmooth problem can be obtained by solving the regularized equation (see [1]) $\sum_{i=1}^{N} \frac{w_i \mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\max\{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|, \delta\}} = \mathbf{0}$ for $\delta > 0$. This is solved using the SOR (Sequential Overrelaxation) algorithm [1] with the overrelaxation parameter $\omega = 1.5$. We choose $\delta = \sqrt{\varepsilon}$ for ε representing the machine precision.

In case of clustering with K prototypes, i.e., the centroids that represent the K clusters, one determines these by solving the nonsmooth problem $\min_{\{\mathbf{c}_k\}_{k=1}^K} \mathcal{J}(\{\mathbf{c}_k\})$, where all $\mathbf{c}_k \in \mathbf{R}^n$ and

$$\mathcal{J}(\{\mathbf{c}_k\}) = \sum_{k=1}^{K} \sum_{i\in I_k} w_i \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{c}_k)\|. \qquad (2)$$

Hereby, I_k determines the subset of data being closest to the kth prototype $\mathbf{c}_k$. The main body of the so-called iterative relocation algorithm for minimizing (2), which is referred as *weighted k-spatialmedians*, consists of successive application of the two main steps: i) find the closest prototype for each observation, and ii) recompute all prototypes $\mathbf{c}_k$ using the attached subset of data. For the latter part, we compute the weighted spatial median as described above. Note that the first step of finding the closest prototype of the ith observation, $\min_k \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{c}_k)\|$, does not need to take the positive weight w_i in (2) into account.

The next issues for the proposed method are the determination of the number of clusters K and the initialization of the clustering algorithm for a given k. Basically, the quality of a cluster can be defined by minimal within-cluster distances and maximal between-cluster distances. Therefore, for the first purpose, we use the approach suggested in [16] and apply three internal cluster indices, namely *Ray-Turi (RT)* [13], *Davies-Bouldin (DB)* [2], and *Davies-Bouldin* (DB*)* [11]. All these indices take both aspects of clustering quality into account: In essence, the clustering error (2), i.e., the sum of the within-cluster distances, to be as small as possible, is divided with the distance between the prototypes (minimum distance for RT and different variants of average distance for DB and DB*), to be as large as possible. When testing a number of possible numbers of prototypes from $k = 2$ into $K_{\max}$, we stop this enlargement when all three cluster indices start to increase.

Concerning the initialization, again partly similarly as in [16], we use a weighted k-means++ algorithm in the initialization of the spatial median based clustering with the weights $\sqrt{w_i}$. A rigorous argument for such an alignment was given in [9] where the relation between variance (weighted k-means) and standard deviation (weighted *k-spatialmedians*) was established. Because of local character, the initialization and the search are repeated $N_s = 10$ times and the solution corresponding to the smallest clustering error in (2) is selected. Furthermore, the weighted k-means++ is applied in the ten initializations with ten different, disjoint data samples (10% of the whole data) that were created using the so-called *Distribution Optimally Balanced, Stratified Folding* as proposed in [12], with the modified implementation given in [7]. Such sampling, by placing a random observation from class j and its $N_s - 1$ nearest class neighbors into different folds, is able to approximate both classwise densities and class frequencies in all the created data samples. Here, we use the 68 country labels as class indicators in stratification.

2.2 Hierarchical application

Because a prototype-based clustering algorithm always works with distances for the whole data, the detection of clusters of different size, especially hierarchically on different scales or levels of abstraction, can be challenging. This is illustrated with the whole PISA data set in Fig. 1, which shows the values of the three cluster indices for $k = 1, \ldots, 68$. For illustration purposes, also the clustering error as defined in (2), denoted as 'Elbow', is provided. All indices have their minimum at $k = 2$ which suggest the division of the PISA data to only two clusters. Note that the geometrical density and low separability of the PISA scale indices might be related to their standardization to have zero mean and unit variance over the OECD countries.

Hierarchical application of the *k-spatialmedians* algorithm was suggested in [17]. The idea is simple: Similarly to the divisive clustering methods, apply the algorithm recursively

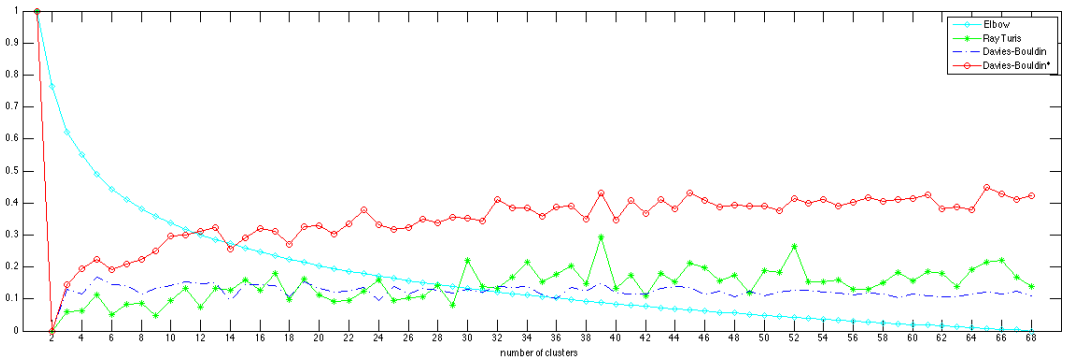

Figure 1: Cluster indices and error slope for the whole sparse PISA data scaled into range $[0, 1]$.

to the cluster data sets that have been determined using the basic approach. For the PISA data here, we realized a recursive search of the *weighted k-spatialmedians* with the depth of three levels, ending up altogether with 2 (level 1), $4 + 4$ (level 2), and $6 + 12 + 10 + 6$ & $2 + 8 + 3 + 6$ clusters (level 3). The wall-clock time for each individual clustering problem was several hours.

3. RESULTS

As discussed in Sec. 1, we use the 15 PISA scale indices that readily summarize most of the students' background as data input for our clustering algorithm. By following the mixture of the partitional/hierarchical clustering approach as described above, we first of all, provide the results of the weighted sparse data clustering algorithm when applied to the whole PISA data (first level). Then, recursively, the results of the algorithm for the newly obtained clusters at the second and third level of refinement are given. For all the clusters at each level, we compute the relative share of students from each country, i.e., the weighted number of students in the cluster in relation to the whole number of 15-year-old students in the country. Moreover, in order to reveal the deviating characteristics of the appearing clusters, we visualize and interpret (i.e., characterize) the cluster prototypes in comparison to the overall behavior of the entire 15-year-old student population in the 68 countries by always subtracting the weighted spatial median of the whole data from the obtained prototypes.

3.1 First Level

Since, as pointed out in Sec. 2.2, all the sparse cluster indices suggest two, we first run our weighted sparse clustering algorithm for $K = 2$. The clustering result on the first level is shown in Fig. 2. The division of these clusters is unambiguous: All scale indices that are associated with high performance in mathematics have a positive value for Cluster 2 and a negative value for Cluster 1. Likewise, those two scale indices that are associated with low performance in mathematics, i.e., the self-responsibility for failing in mathematics (*failmat*) and the anxiety towards mathematics (*anxmat*), show a positive value for Cluster 1 and a negative value for Cluster 2. As can be expected by these profiles, the mean mathematics performance of Cluster 1 is much lower than the mean performance of Cluster 2 (see Table 1).

When we consider the relative number of students from different countries, we see that every country has students in both clusters. In fact, the distribution of the 15-year-old student population between the two clusters is quite equal in each country. For Cluster 1, the mean percentage of students from a country is 55% while for Cluster 2, the mean is 45%, and both have the standard deviation of 10. In all of the in PISA participating countries and territories, there are higher and lower performing students and it seems that they share the same characteristics. Additionally, the distribution between girls and boys is quite equal, although somewhat in favor of boys: Only 48% of the students in the cluster with the scale indices that are associated with high performance in mathematics are girls. Moreover, the average math score of the boys is in both clusters higher than the average math score of the girls (see Table 1).

Table 1: Characteristics of global/first level clusters.

Cluster	**population size** (♀ in %)	**math score** ∅	♀	♂
1	13399687 (52%)	445	442	449
2	11321033 (48%)	468	461	475
all	24720720 (50%)	456	451	461

3.2 Second Level

Following the approach as described above, we run the clustering algorithm again, but this time for each of the two global clusters obtained in the first level separately. According to the same rule given in Sec. 2.1, i.e., stop enlarging k during the search when all the cluster indices are increasing, we get for both of the global clusters $K = 4$ as a number for their subclusters.

3.2.1 Subclusters of Cluster 1

Table 2: Characteristics of subclusters of Cluster 1.

Cluster	**population size** (♀ in %)	**math score** ∅	♀	♂
1-1	2792046 (56%)	439	438	440
1-2	3873035 (52%)	391	388	394
1-3	3072064 (58%)	466	464	468
1-4	3662542 (45%)	491	489	492

The subclusters of the global Cluster 1 are visualized in Fig. 3 and characterized in Table 2. If we set the threshold

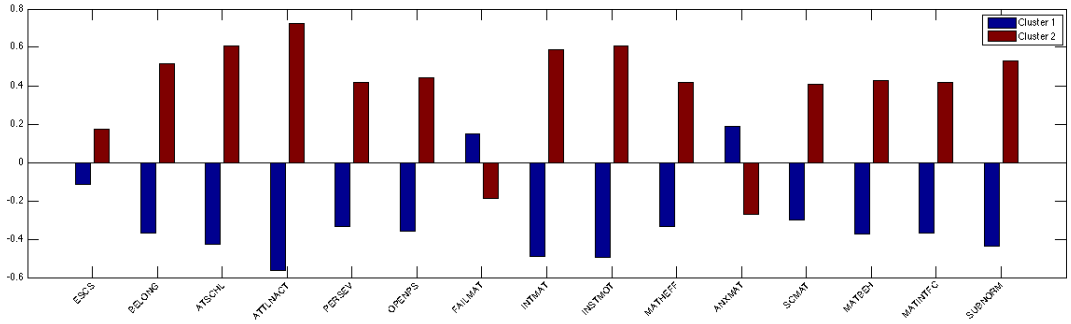

Figure 2: Characterization of the two global clusters.

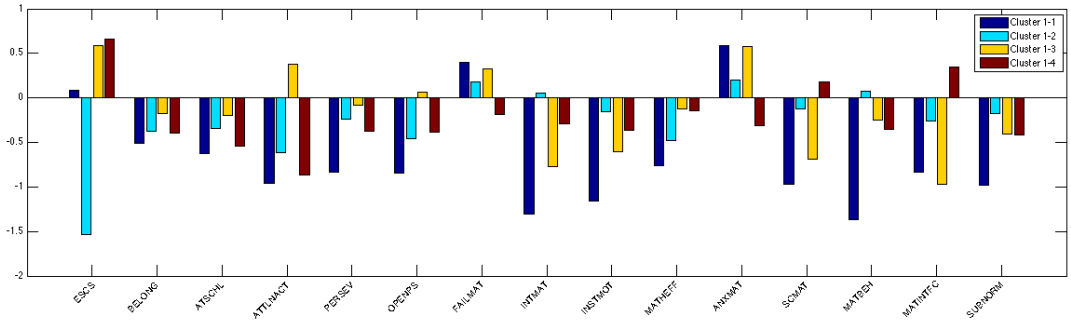

Figure 3: Characterization of the four subclusters of Cluster 1.

of how many students should at least be from one country to 21%, we obtain the following countries for the subclusters: Cluster 1-1 (i.e., subcluster 1 of Cluster 1) contains at most students from East Asia with the exception of China: More than 30% of Japan's 15-year-old student population *belong*s to this cluster, 26% of Korea's and and 25% of Taiwan's. The remaining students represent a mixture from many different countries which, however, are only represented by less 21% of their 15-year-old student population.

Cluster 1-2 contains almost entirely students from developing countries. Hereby, students from Vietnam form with 49% the majority. Moreover, Indonesia, Thailand (both $> 30\%$) and Brazil, Colombia, Peru, Tunisia, and Turkey (all $> 25\%$) are represented by this cluster. The cluster is, as can be seen from Fig. 3, most notably characterized by a very low economic, social and cultural status (*escs*). That means that the students in this cluster - as a subset of the global Cluster 1 which already represented the more disadvantaged students (see Fig. 2) - are the most disadvantaged.

Cluster 1-3 consists in the majority of students from Eastern Europe: Serbia, Montenegro, Hungary, Slovak Republic (all $> 23\%$) and Romania (almost 22%) constitute the majority. As we can see from Fig. 3, this cluster is the only one in the group of subclusters of the global Cluster 1, that generally was characterized by negative attitudes and perceptions (see Fig. 2), which actually can be distinguished by positive attitudes towards school (*attlnact*). Moreover, it is the cluster with mainly girls in it.

Cluster 1-4 accommodates mainly students from Western and Central Europe. Most of the 15-year-old student population from the Netherlands (39%) are in this cluster, followed by Belgium with 29%, and the Czech Republic with 27%. This cluster is characterized by the highest *escs* among the students of the global Cluster 1. Furthermore, although they have negative values in most of the scale indices, they have a higher mathematics self-concept, and also much higher intentions to use mathematics later in life in comparison with their peers.

3.2.2 *Subclusters of global Cluster 2*

Table 3: Characteristics of subclusters of Cluster 2.

Cluster	**population size** (♀ in %)	**math score** ∅	♀	♂
2-1	3127958 (43%)	526	523	528
2-2	2739481 (54%)	457	457	458
2-3	3521092 (50%)	400	397	403
2-4	1932502 (44%)	515	506	523

The subclusters of the global Cluster 2 are characterized in Fig. 4 and summarized in Table 3. Again, we search for clusters that mostly deviate from the others. Cluster 2-1 is such a cluster: The students in this cluster have the highest average math score (see Table 3), the highest intentions to pursue a mathematics related career but a sense of *belong*ing to school (*belong*) and subjective norms in mathematics (*subnorm*) that are only about the same as the average of the whole 15-year-old student population (see Fig. 4). The subjective norms in mathematics measure how people important to the students, such as their friends and parents, view mathematics. In the global Cluster 2, those students

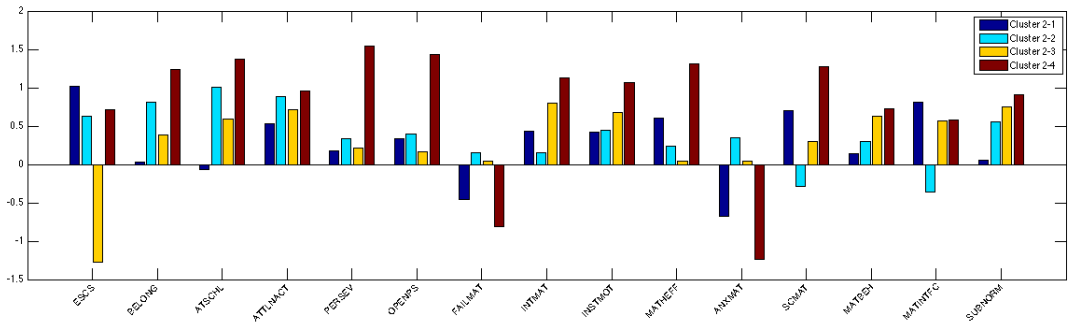

Figure 4: Characterization of subclusters of Cluster 2.

who had high positive values in the other scale indices associated with high performance in mathematics, also thought that their friends and family view mathematics as important (their *subnorm* value is very high, see Fig. 2). Students in this cluster, however, seem not to be influenced or affected by what people close to them think. It appears to be a rather strong cluster that also has the highest percentage of boys in it. For this cluster, we again compute the relative number of students from each country. And indeed, it shows a very clear country-profile. The highest percentage of students come from the English-speaking and Nordic countries: Denmark (more than 30%), Iceland and Sweden (both $> 26\%$) have the highest percentages of their 15-year-old student population in this cluster. Followed by the two highest performing districts in the USA, namely Connecticut and Massachusetts, with both more than 25%. Besides these countries and territories, the cluster has also a high share of students from Norway, Finland, Great Britain, Australia, and Canada (almost 22% or more). Additionally, the USA has with more than 21% still a relatively high share of students in this cluster. According to the Hofstede Model (see Sec. 1), all of these countries are characterized by high individualism.

Also Cluster 2-3 shows an explicit country profile: 36% of the 15-year-old student population from India are in this cluster. Moreover, the cluster consists of students from Peru and Thailand (both 30%), Turkey (27%) and Vietnam (26%). Altogether, we find here the most disadvantaged students (indicated by the very negative *escs*) among the subgroups of the global Cluster 2 and the largest share of students come from the developing countries. However, these students have very positive attitudes towards education and show relatively high values in all scale indices that are associated with high performance in mathematics.

To this end, Cluster 2-2 and Cluster 2-4 have less obvious country affiliations. Cluster 2-2 can at best be described as containing mostly countries with Islamic culture. Most of the students are from the United Arab Emirates and Albania (both 21%), Kazakhstan and Jordan (both 19%). According to the Hofstede Model, these countries are similar in that way that they all show very high power distance. Cluster 2-4 has with 25% the highest share of students also from Kazakhstan, but the remaining countries in this cluster (all have less than 17% of their 15-year-old students population in it) are widely mixed.

Altogether, among the clusters at the second level, Cluster 2-1 appears to be the most interesting one, i.e., the most distinct group with the clearest country profiles.

3.3 Third Level

Recursively, we repeat the same approach on the next level, i.e., for the subclusters of the eight clusters identified in Sec. 3.2. For all the new subclusters, the best number of clusters as determined by the cluster indices are as follows: 6, 12, 10, and 6 for the four subclusters of the first global cluster, and 2, 8, 3, and 6 for the four subclusters of the second global cluster. This means that we have 53 different clusters on this level - almost as many as different countries/territories in the whole PISA 2012 data. If our hypothesis is true, we should be able to find clusters that clearly contain more students from certain countries. Exactly as in Sec. 3.2, we first of all compute the basic facts of each cluster and characterize the prototype that describes the profile of the particular cluster.

3.3.1 Subclusters of Cluster 1-3

Table 4: Characteristics of subclusters of Cluster 1-3

Cluster	**population size** (♀ in %)	**math score** ∅	♀	♂
1-3-1	335240 (61%)	493	492	495
1-3-2	262779 (48%)	539	540	538
1-3-3	368591 (51%)	461	460	462
1-3-4	273629 (66%)	492	491	492
1-3-5	359721(56%)	427	428	426
1-3-6	275513 (63%)	437	436	438
1-3-7	264017 (63%)	443	441	447
1-3-8	318607 (63%)	460	457	464
1-3-9	216704 (60%)	421	418	424
1-3-10	397263 (56%)	481	482	480

The first interesting cluster appears in the 1-3 group. Cluster 1-3-8 accommodates mainly students from South West Europe: Austria, Liechtenstein, Spain, France, and Italy. According to the Hofstede Model, all of these countries are depicted by high avoidance of uncertainty.

3.3.2 Subclusters of Cluster 1-4

The characterization of the subclusters in the 1-4 group are provided in Fig. 6, and summarized in Table 5. Also here, we are searching for explicit country clusters. This search is realized by looking at the histograms and identifying those

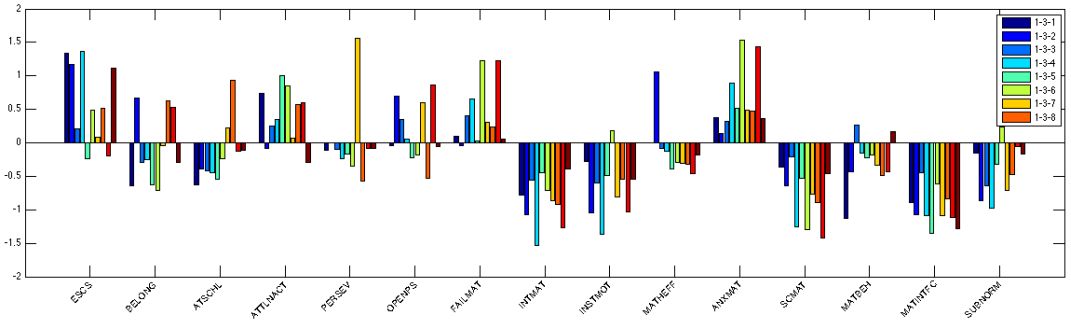

Figure 5: Characterization of subclusters of Cluster 1-3.

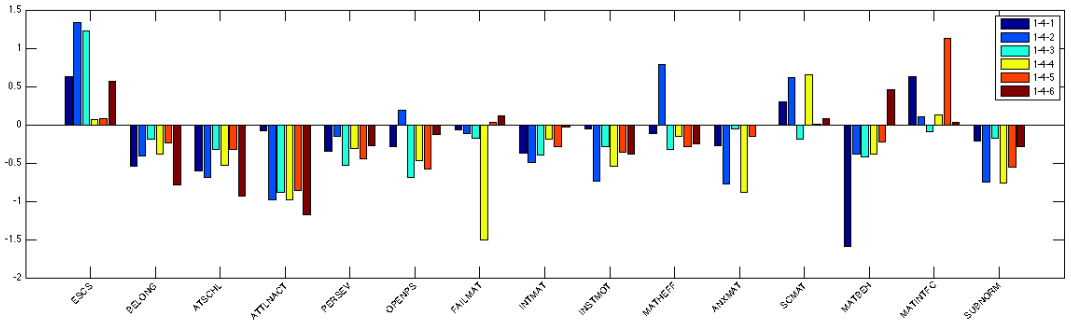

Figure 6: Characterization of the subclusters of Cluster 1-4.

Table 5: Characteristics of subclusters of Cluster 1-4

Cluster	population size (♀ in %)	math score ∅	♀	♂
1-4-1	485599 (48%)	481	480	482
1-4-2	520763 (38%)	556	558	555
1-4-3	771799 (53%)	494	494	495
1-4-4	489528 (43%)	497	491	501
1-4-5	754515 (48%)	470	467	473
1-4-6	640338 (38%)	461	465	458’

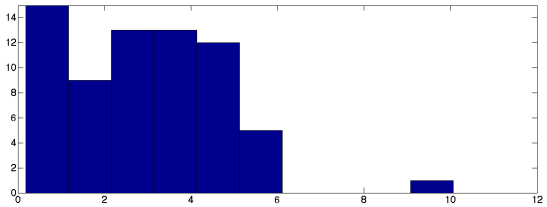

Figure 7: Histogram of the distribution of countries from the students in Cluster 1-4-2.

clusters that for some countries have a considerably higher share of their 15-year-old student population in it than for the remaining countries. The histogram in Fig. 7 shows one example of this for Cluster 1-4-2: In this cluster, the portion of students in it deviates significantly from the others for exactly one country with 10% of its 15-year-old student population. This country is the Netherlands. For all other countries, the share of their 15-year-old student population in this cluster is less than 6% (see Fig. 7). As can be seen from Fig. 6, this ‘Netherlands Cluster’ is characterized by having the highest math self-efficacy amongst its group.

Cluster 1-4-1 is again a mixture of Nordic and English-speaking countries. The highest share of students in this cluster come from the United Kingdom, Ireland, Norway, New Zealand, and Sweden. As these two country profiles were already detected to be in the same cluster on the higher cluster level (see Sec. 3.2.1), it really seems that students from these countries share many similar characteristics.

Cluster 1-4-4 has the highest share of East Asian countries including two of the three districts of China that participated in PISA 2012. Most of the students in this cluster come from Japan, followed by Taiwan, Macao-China and Hong Kong-China. One of the most distinct feature of this cluster is, as can be seen from Fig. 6, the high self-concept in mathematics (*scmat*). According to the Hofstede Model (see Sec. 1), all of these countries show high pragmatism.

3.3.3 *Subclusters of Cluster 2-1*

Table 6: Characteristics of subclusters of Cluster 2-1

Cluster	population size (♀ in %)	math score ∅	♀	♂
2-1-1	1346930 (40%)	562	557	566
2-1-2	1781028 (45%)	498	500	497

From Sec. 3.2, we concluded that Cluster 2-1 was the most interesting one. Moreover, Cluster 2-1 was the cluster that had the highest share of two country profiles in it: On the one hand, the English-speaking countries, and, on the other hand, the Nordic countries. Interestingly, the cluster indices

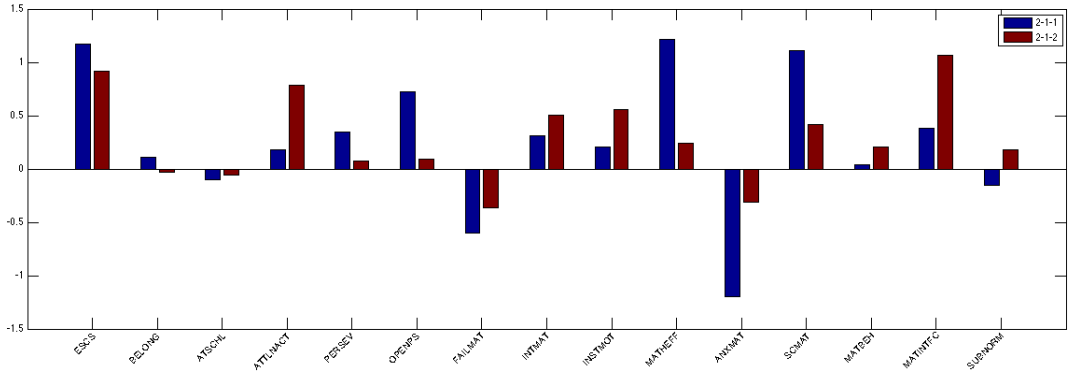

Figure 8: Characterization of subclusters of Cluster 2-1.

also suggest to divide this cluster into two further countries. However, when we look again at those countries that have the highest percentages of their 15-year-old students, the two clusters still contain mostly students from both country profiles. For example, 15% of the Danish 15-year-old student population are in Cluster 2-1-1, and 14% are in Cluster 2-1-2. Similarly, 14% of the 15-year-old student population from Connecticut are in Cluster 2-1-1, and 11% in Cluster 2-1-2. Apparently, this cluster does not divide any further between Nordic and English-speaking countries. It only divides the high-performing students from these countries into two types: On the one hand, the type that has a very high self-efficacy (*matheff*) as well as self-concept (*scmat*) in mathematics, i.e., the students that have a very high belief in their own ability, and, on the other hand, the type that has very high intentions to pursue a mathematics related career (*matintfc*).

However, also a new clear group of countries appears. Cluster 2-1-1 has a very high share of German-speaking countries in it: More than 12% of Germany's and Switzerland's 15-year-old student population, and 10% of Austria's can be found in this cluster. None of these countries appear in the sibling Cluster 2-1-2 when the threshold is set to 9%. It seems that German-speaking students feel very confident of solving mathematical tasks but only show a moderate positive value in the intentions to use mathematics later in life, a characteristic that one would associate the most with the traditional functional German stereotype (see Sec. 1) that is expected to attach great importance to utilitarianism [4]. According to the Hofstede Model, all of these three German-speaking countries are considered to be highly masculine.

3.3.4 *Subclusters of Cluster 2-4*

Table 7: Characteristics of subclusters of Cluster 2-4

Cluster	**population size** (♀ in %)	**math score** ∅	♀	♂
2-4-1	186107 (37%)	533	528	536
2-4-2	430729 (40%)	582	575	588
2-4-3	261838 (45%)	440	436	443
2-4-4	378120 (50%)	477	468	486
2-4-5	430105 (47%)	520	519	521
2-4-6	245603 (40%)	516	500	526

The subclusters of Cluster 2-4 are summarized in Table 7 and characterized in Fig. 9. The clearest country profile among this group is 2-4-6: It consists to the highest share of students from high-performing Asian countries: Shanghai-China and Singapore. As we can see from Fig. 9, similarly to Cluster 1-4-4 (see Sec. 3.3.2) that also contained a high share of Chinese students, this cluster is characterized as well by a high self-concept in mathematics (*scmat*). The students in this cluster believe that mathematics is one of their best subjects, and that they understand even the most difficult work. Furthermore, as already found for Cluster 1-4-4, also for this cluster the main countries show high pragmatism according to the Hofstede Model.

4. CONCLUSIONS

In this article, we have introduced a clustering approach that has both partitional and hierarchical components in it. Moreover, the algorithm takes weights, aligning a sample with its population into account and is suitable for large data sets in which many missing values are present.

The hypothesis in our study was that the different clusters determined by the algorithm, when all students with their attitudes and behaviors towards education are given as input, could be explained by the country of the students in particular clusters. Our overall results on the first level showed that in each cluster students from all countries exist and that the actual test performance (as well as a simple division in positive and negative attitudes towards education) explain the clusters much better than the country from which the students in the particular cluster come from.

However, on the next two levels many clusters were detected that obviously had a much higher share from students from certain countries. For example, an Eastern Europe, a German-speaking, an East Asia, and a developing countries cluster were identified. On the second level, also a very clear cluster that consisted to a high portion of Nordic and English-speaking countries appeared. This cluster did not split further on the next level to fully separate these two distinct country profiles. Instead, the cluster was divided into two student types, of which both the Nordic as well as the English-speaking countries seem to have an almost equal share of their students from.

Summing up, we conclude that groups of similar countries, e.g., by means of geographical location, culture, stage of de-

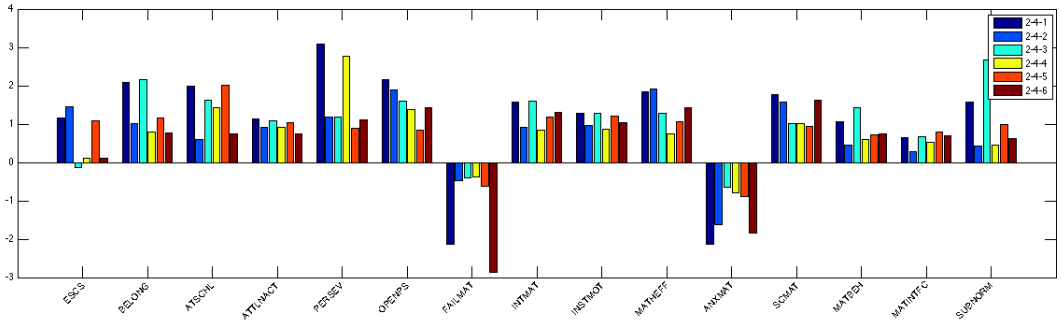

Figure 9: Characterization of subclusters of Cluster 2-4.

velopment, and dimensions according to the Hofstede Model, can be found by clustering PISA scale indices but the actual country stereotypes exist only to a very marginal extent. However, in a further work it would be interesting to include more variables to the algorithm than the 15 scale indices utilized here. The PISA scale indices are linked to the performance in mathematics and in every country there are higher and lower performing students who share similar overall characteristics.

The overall results presented here show a very promising behavior already, and we expect that the resulting clusters of our algorithm could be explained even clearer by the country, if information such as the students' temperament would be available for the clustering algorithm. However, from the methodological perspective, one faces difficulties in establishing clear rules and thresholds to distinguish significant findings and characterizations of clusters, e.g., by means of their relative share of students from a certain country.

References

[1] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[2] D. L. Davies and D. W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.

[3] S. Harrell. Why do the Chinese work so hard? Reflections on an entrepreneurial ethic. *Modern China*, pages 203–226, 1985.

[4] M. F. Herz and A. Diamantopoulos. Activation of country stereotypes: automaticity, consonance, and impact. *Journal of the Academy of Marketing Science*, 41(4):400–417, 2013.

[5] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Edward Arnold, London, 1998.

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

[7] T. Kärkkäinen. On cross-validation for MLP model evaluation. In *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science (8621), pages 291–300. Springer-Verlag, 2014.

[8] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[9] T. Kärkkäinen and M. Saarela. Robust principal component analysis of data with missing values. *To appear in the Proceedings of the 11th International Conference on Machine Learning and Data Mining MLDM*, 2015.

[10] T. Kärkkäinen and J. Toivanen. Building blocks for odd-even multigrid with applications to reduced systems. *Journal of Computational and Applied Mathematics*, 131:15–33, 2001.

[11] M. Kim and R. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[12] J. Moreno-Torres, J. Sáez, and F. Herrera. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.

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

[14] M. Saarela and T. Kärkkäinen. Discovering Gender-Specific Knowledge from Finnish Basic Education using PISA Scale Indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[15] M. Saarela and T. Kärkkäinen. Analysing Student Performance using Sparse Data of Core Bachelor Courses. *JEDM-Journal of Educational Data Mining*, 7(1):3–32, 2015.

[16] M. Saarela and T. Kärkkäinen. Weighted clustering of sparse educational data. *To appear in 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[17] P. Wartiainen and T. Kärkkäinen. Hierarchical, prototype-based clustering of multiple time series with missing values. *To appear in 23rd Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.