CLUSTERING EU COUNTRIES BASED ON DEATH PROBABILITIES

Kolos Csaba Ágoston
Institute of Mathematics and Statistical Modelling
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
Centre for Economic and Regional Studies
Tóth Kálmán u. 4. Budapest 1097, Hungary
E-mail: kolos.agoston@uni-corvinus.hu

Ágnes Vaskövi
Institute of Finance, Accounting and Business Law
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: agnes.vaskovi@uni-corvinus.hu

KEYWORDS

Mortality, Clustering, Death Probabilities

ABSTRACT

Background Our research is conducted to identify certain grouping of 24 European countries based on their death probabilities. Gathering 2014 data from Human Mortality Database our research *objective* was twofold. First, we wanted to find homogeneous groups of countries where mortality is similar and for a financial institution they could be grouped as risk communities. Second, we wanted to identify the optimal number of groups as a basis for strategy making. Two different clustering *methods* were used in our research, k-means and k-median clustering. We applied asymmetric measure (QDEV) in k-median method to handle the differences in country sizes and age groups. Our *results* are stable but different in k=3 clusters, k-means clustering resulted in a big Western-European cluster and two small-medium Eastern groups; however, k-median clustering gave a homogeneous Eastern group and besides a bigger Western cluster Spain, Italy, and France formed a separated group of countries.

INTRODUCTION

In our globalizing world it is highly important for all industries to shift from individual customer service to grouped solutions. Defining homogeneous groups of customers decreases production and service costs of companies and it could create value for the customers at the same time. Global insurance companies and financial institutions active in several countries might be aware of life expectancy and death probability differences across countries in order to decrease their longevity risk. Nevertheless, insurance companies place great emphasis on the establishment of homogeneous risk communities. The heterogeneity of the insureds is well known and is further amplified by the phenomenon of anti-selection. Insurance companies aim to establish homogeneous risk communities, however in terms of the price calculation it is not favorable if some risk communities get too fragmented. The allocation of the heterogeneous insureds into (somewhat) homogeneous groups is called risk classification in the literature (see Crocker and Snow (1986)). Risk classification might be the most often used actuarial method that can be supported by the adequately chosen cluster analytical methods.

In the case of life insurances, some factors of heterogeneity have been known for centuries, for example the difference between the male and female death probabilities. In the last decades deeper analyses had also been revealed – thanks to the evolution of computing technologies. The difference between the mortality pattern of the white and blue collars became general knowledge, and the difference in death probabilities based on educational attainment is increasingly recognized. The territorial diversities are also more and more obvious. Kovács and Vaskövi (2019) also used cluster analysis to group European countries base on their life expectancy and retirement age patterns. In this paper we investigate the national differences of unisex death probabilities in 24 European countries based on their 2014 data and give possible classification using different clustering methods. We include death probabilities separately for former East and West-Germany to identify possible remaining differences.

DATA: DEATH PROBABILITIES OF EU COUNTRIES

Individual longevity is uncertain, thus the length of human lifetime can be described by a random variable. Although there were specific attempts to describe the distribution of this variable with a functional form (see Marshall and Olkin (2007)), sufficient result is still missing. Instead, age specific death probabilities are calculated: q_x gives the probabilities that a living x-year-old person will die within a year. These rates can be calculated based on institutional (insurance companies or pension funds) data or based on nationwide statistics. The estimation process differs in some extent for the two cases: for institutional data mostly the Kaplan-Meier method is used; however, for nationwide statistics the so called Lexis-diagram is used. Both methods belong to nonparametric statistical methods, i.e. neither Kaplan-Meier method nor Lexis diagram assumes data fits normal or any well-understood distributions.

We used unisex crude death probabilities (a certain averaging is used to calculate unisex rates from male and female death probabilities) for EU countries available in Human Mortality Database (HMD). In Table 1 unisex death probabilities in year 2014 of 3 chosen countries

Communications of the ECMS, Volume 34, Issue 1,
Proceedings, ©ECMS Mike Steglich, Christian Mueller,
Gaby Neumann, Mathias Walther (Editors)
ISBN: 978-3-937436-68-5/978-3-937436-69-2(CD) ISSN 2522-2414

are shown in every 5 years (former East and West-Germany are described separately):

Table 1: Death probabilities (q_x) of 3 European countries from age 1 to 110 in 2014

years	AUT	DE-E	DE-W	HUN
1	0.00019	0.00028	0.00024	0.00036
5	0.00009	0.00007	0.00006	0.00011
10	0.00009	0.00007	0.00005	0.00011
20	0.00038	0.0004	0.0003	0.00038
30	0.00042	0.00049	0.00041	0.00051
40	0.00096	0.00105	0.00085	0.00161
50	0.00257	**0.00345**	0.00266	0.00588
60	0.00699	0.00828	0.0074	0.01485
70	0.01626	0.01667	0.01682	0.02904
80	0.04379	0.04816	0.04536	0.06906
90	0.14823	**0.1515**	0.15042	0.17988
100	0.3746	0.3684	0.37224	0.37025
110	1.0000	1.0000	1.0000	1.0000

For example, bold figures in Table 1 mean the probability that an East-German individual at the age of 50 dies within one year is 0.35% and at the age of 90 is 15.15%. Crude death probabilities are significantly different in ages, moreover for smaller countries we face particular ages without deaths. For this reason, the crude death probabilities are smoothed (Ágoston, 2003). Smoothing method differ from country to country, for child ages and for young adults a polynomial function (with high degree) is fitted or some kind of moving average method is applied. For adults and old ages crude probabilities often smoothed based on Gompertz-Makeham (Gompertz, 1825 and Makeham, 1867) mortality law.

Standardized data is used in order to reduce the dispersion of elderly death probabilities since the latter can be a magnitude higher than the ones in younger ages. If we did not standardize data, the clusters would only be formed based on the elderly mortality.

Infant death probabilities are omitted from the database considering three main reasons, (i) infant mortality (probability of death in the age group 0 to 1) is significantly higher than it is in other child age groups, (ii) the intrauterine death is not clearly classified, (iii) the death probability of children between 0 to 1 years has changed significantly in the last decades and this change was not in line with the change of any other age groups.

In the next section we describe the clustering methodologies used to classify EU countries based on their death probabilities. Vékás (2019) found empirical evidence of changing mortality curves in European countries, and in our paper we also attempt to identify mortality patterns among the examined countries.

METHODOLOGY

Cluster analytical methods on mortality data appear in Ágoston, Majstorović and Vaskövi (2019). They used parametric methods where first the Makeham mortality law (Makeham, 1867) is fitted to the data then the clustering method is applied on the fitted survival curves. The Makeham mortality law fits well on data of age groups 30 to 100 years; however, in this paper we wanted to analyze also death probabilities of younger ages, thus crude death probabilities were clustered here. K-means and k-median clustering methods were applied on data and in k-median method we used an asymmetric similarity measure (*QDEV*) described by Arató et al. (2009).

K-means Method

One of the first, and most popular clustering method until today is the k-means method (McQueen, 1967). The method can classify the observations in such a way, that the squared sum of distances in the groups are minimal. The method is very fast but highly dependent on the selection of initial cluster center (it is a computed center not a real data), thus it can be viewed as a heuristic approach.

If we use clustering method based on death probabilities the most straightforward way is that we take the death probabilities for ages 1..*N*, we consider it as an *N* dimensional vector in the *N* dimensional space and use a k-means algorithm for these vectors.

Using k-means method when (re)calculated the cluster centers, the *mean* of death probabilities is calculated. It can happen that we simply take the average of Germany and Luxemburg, although German data is based on 80.77 million people and Luxembourgian data is based on 0.55 million people. In these specific cases where the size of countries is very much different, the mean would be problematic. But we can move further: when we calculate the distance between cluster center and instances we use Euclidean distance, i.e. all ages play the same important role although the sample size can differ greatly from ages to ages; old age probabilities are calculated based on significantly less data than middle age probabilities.

There would be a straightforward option to give weights for ages. The problem is if the sample size (even its relative value) differs for countries, it would not be fortunate to use the same weights for a given age for all countries. Another approach is to specify a distance or similarity measure that can consider the number of entities in an age group. Arató et al. (2009) define three different similarity measures for life tables, for our purpose the *QDEV* measure is relevant. We consider two countries (*a* and *b*) and the similarity between them is defined by:

$$QDEV = \sum_{i=0}^{100} \frac{e_i^a (q_i^a - q_i^b)^2}{q_i^b} \qquad (1)$$

where e is the so called exposure (the time while an x-year-old person was alive, often quite close to the number of individuals alive). In expression (1) the term $(q_i^a - q_i^b)^2$ is squared therefore the major difference between death probabilities has significantly higher importance than many small differences together. The previous term is normalized by probability of country b, meaning greater difference can be tolerated if the probability itself is a higher value, and multiplied by the exposure, meaning that for small communities even high differences can be tolerated.

We can see that expression (1) is not symmetric in a and b. In cluster analysis similarity measures are usually symmetric but we can find examples for asymmetric measures, as well (see Okada, 2000). We have two options: keep the measure asymmetric and chose a method which can handle asymmetric measure or symmetrize it somehow (for instance the average of the two direction). We tried both ways but in calculations of this paper kept the measure asymmetric.

For asymmetric measure it is not straightforward how cluster centers should be calculated in k-means method (even if we symmetrize the QDEV measure, it is still problematic). Although Olszewski (Olszewski 2011) gives a modified k-means algorithm for asymmetric measures, the algorithm was not tested on real data; therefore, we suggest to use k-median clustering method that is suitable for the asymmetric QDEV similarity measure.

K-median Method

The so-called k-median (or also p-median) problem was already researched at the dawn of the appearance of cluster analytical methods. The idea was, that if we minimize the absolute deviation instead of the squared sum, then in case of a one-dimensional problem the cluster centers will be data points (or at least can be chosen to be data point). Unlike the average, the median is not defined in multi-dimensional space but the name stayed with the method. In case of a multi-dimensional k-median problem a distance or similarity matrix calculated first, and the task is to decide which data points should be cluster centers and which cluster center and data point should be matched to in order to minimize the total distance from the centers.

We calculated the QDEV equation for every possible country pairs and a similarity matrix is produced. This matrix is the input for the k-median method.

RESULTS

Results of K-means Method

If the number of clusters is 2, we get reasonable but trivial groups: Central-European and Baltic countries form the first cluster, Western countries forms the second. k=2 clustering is stable, while all 10,000 runs with random initial cluster center gave the same result of 1326.15 between groups distance.

Table 2: k-means clustering with 2 clusters

Cluster Id	Countries in the cluster
1	Czech Republic, Slovakia, Hungary, Poland, Estonia, Lithuania, Latvia, Bulgaria, Croatia,
2	Austria, East-Germany, West-Germany, Belgium, Netherlands, Luxemburg, France, Great-Britain, Ireland Denmark, Finland, Sweden, Portugal, Spain, Italy, Slovenia

Slovenia belongs to the second cluster together with all developed countries; however, the three Baltic countries are grouped to the eastern block. This clustering suggests that rapid economic development of Baltic countries was not accompanied by a significant improvement of demographic processes. Based on clustering of 2014 death probabilities we did not find difference between East and West-Germany.

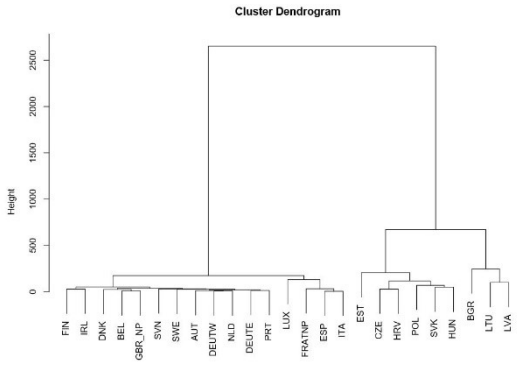
When we raised the number of clusters up to 3 then the first cluster is divided into two separated groups where Bulgaria, Lithuania and Latvia form one cluster and the remaining 6 countries from the first cluster form the second. Clusters are still very stable (6,017 out of 10,000 runs come to the same between groups distance of 1,661.922).

We raised the number of clusters up to k=6 to gain more detailed grouping. In Table 3 clusters of countries are shown. Raising the number of clusters up to 4 Bulgaria would form an individual group; then Estonia leaves its cluster. The biggest cluster of Western countries comes apart only at k=6 where Spain, Italy, France, and Luxemburg form the sixth cluster. East and West-Germany remain in the same cluster meaning there is no significant difference between death probabilities of the two parts of the country.

Table 3: k-means clustering with 4, 5 and 6 clusters

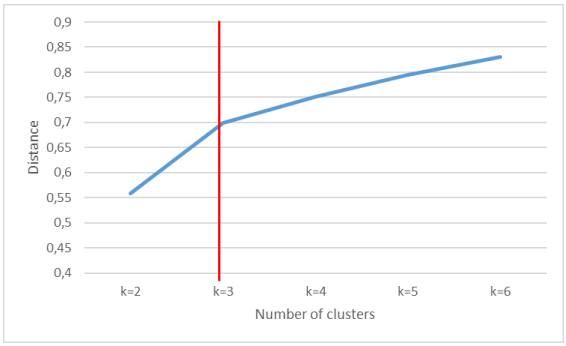
Cluster Id	k=4	k=5	k=6
1	BGR	BGR	BGR
2	LTU, LVA	LTU, LVA	LTU, LVA
3	CZE, EST, SVK, HRV, HUN, POL	CZE, SVK, HRV, HUN, POL	CZE, SVK, HRV, HUN, POL
4	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, ESP, ITA, LUX, FRA	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, ESP, ITA, LUX, FRA	AUT, DEUE, DEUW, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT
5		EST	EST
6			ESP, ITA, LUX, FRA

Since similarity matrix is applied in k-means clustering we could also assign hierarchical clustering method on our data. Figures 1 shows the clustering steps and distances of the 24 countries visualizing the results of the k=6 k-means clustering by a dendrogram (Ward method). The 12 West-European countries forming cluster No 4 are the most similar with the minimum dissimilarity measure. Cluster No 6 is formed by 4 countries (Luxemburg, Italy, Spain, and France) joining cluster No 4 in the second step. The East-European countries form two main groups (cluster No 2 and 3); however, Estonia and Bulgaria have significantly different death probabilities both forming one distinct group (cluster No 1 and 5).



Figures 1: Dendrogram of hierarchical clustering, Ward method

We also applied cluster elbow method which is a heuristic way to define the optimal number of clusters. The variance explained is calculated as a ratio of between-group-VAR and total-VAR and this percentage is plotted. Increasing the number of clusters would raise the variance explained; however, the marginal gain is flattening. Where the slope of the curve decreases there is the "cluster elbow", i.e. the optimal number of clusters. Figures 2 shows the cluster elbow of k-means clustering equals to 3.



Figures 2: Cluster elbow diagram of k-means clustering

Over k=4 the k-means clustering could not be considered stable since the random cluster centers result in wide variety of between groups distances.

Results of K-median Method with QDEV Similarity Measure

In k-median method the cluster centers are real data points, in our case they are countries that the most typical countries are in the cluster.

When k=2 the same clusters were produced as with k-means method, i.e. blocks of 15 Western and 9 Eastern European countries. Belgium is the cluster center (the most typical country) in West-Europe and Slovakia in East-Europe.

In case of k=3 the Western-block is divided into two smaller groups where Spain, Italy and France formed a new group of countries. The same clustering result is to be identified at higher number of clusters in k-means method (k=6). Belgium as cluster center in West-Europe is replaced by (former) West-Germany, and in the new cluster Spain is pointed as cluster center. Slovakia remained the center of East-European countries. From this cluster number, we could observe that big countries become cluster centers.

In Table 4 results of $4 \leq k \leq 6$ clustering are summarized. Countries indicated bold and underlined are the cluster centers of each group. At k=4 the eastern-block is divided; Hungary, Bulgaria, Lithuania, and Latvia form cluster No 1. In k-median clustering, Hungary does not connected to other Visegrad countries but to Bulgaria and the Baltic countries. When we further increased the number of clusters, Eastern blocks were not changed, but France and Sweden were moved from their cluster. The former East and West-Germany remain in the same cluster meaning there is no difference of death probabilities in the two halves of the German country.

Table 4: k-median clustering for $3 < k \leq 6$ clusters

Cluster Id	k=4	k=5	k=6
1	HUN, **BGR** LTU, LVA	HUN, **BGR** LTU, LVA	HUN, **BGR** LTU, LVA
2	CZE, **POL**, EST, SVK, HRV	CZE, **POL**, EST, SVK, HRV	CZE, **POL**, EST, SVK, HRV
3	AUT, DEUE, **DEUW**, BEL, NLD, DNK, FIN, SWE, GBR, IRL, SVN, PRT, LUX	AUT, DEUE, **DEUW**, BEL, NLD, DNK, FIN, GBR, IRL, SVN, PRT, LUX	AUT, DEUE, **DEUW**, BEL, NLD, DNK PRT, SVN,
4	**ESP**, ITA, FRA	ESP, **ITA**, SWE	ESP, **ITA**, SWE
5		**FRA**	**FRA**
6			FIN, LUX, **GBR**, IRL

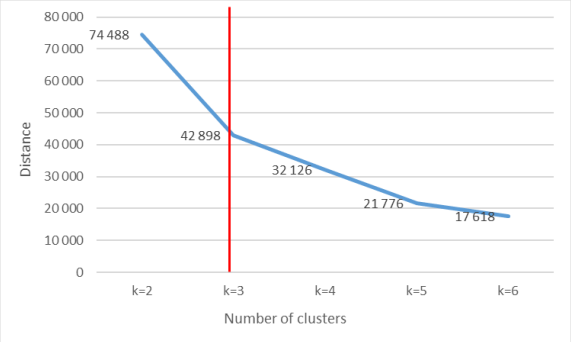
In the case of a k-median problem (or also by k-means problem), the number of clusters is an input parameter; therefore, selection of the optimal cluster number must

be part of the analysis. In k-median method the objective function is calculated which is the sum of within-group distance measures, shown in Table 5.

Table 5: values of objective functions at k-median

	Objective functions
k = 2	74,487.61
k = 3	42,897.50
k = 4	32,125.75
k = 5	21,775.68
k = 6	17,617.9

Figures 3 shows the optimal cluster number of k-median clustering using the values of objective functions.

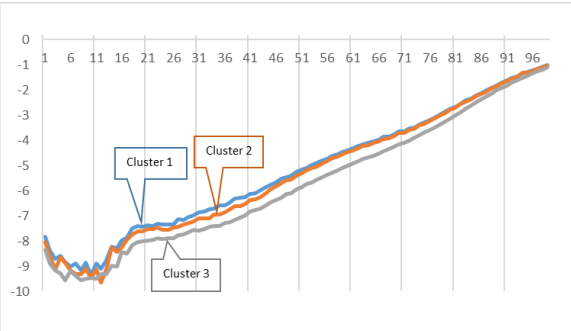


Figures 3: Cluster elbow diagram with objective functions of k-median clustering

Comparison of K-means and K-median Results

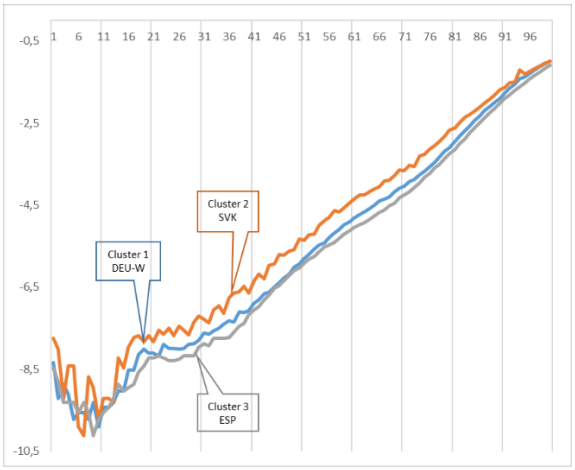
In k-median method the cluster centers are real data points, in our case these are the most typical countries of each cluster. On the contrary, cluster centers in k-means method are computed centers and very rare coincide with certain data point.

Figures 3 shows the logarithmic death probabilities of k-means clusters (k=3) on a log scale. Cluster 1 (Bulgaria, Lithuania and Latvia) has the highest death probabilities in most of the age groups; however, from age group 70 the death probability curves of Cluster 1 and 2 overlap. Cluster 3 (15 Western-European countries) has the lowest age-specific probability of death for the total 1-100 years. There is one age group (12-13 years) where the probability of death is lower in cluster 2 (Eastern countries) than is cluster 3.



Figures 4: logarithmic death probabilities for age groups, k-means clustering (k=3)

We explained in the previous section that using k-median clustering method the optimal number of clusters is again three. Thus, on Figures 5 we represent logarithmic death probabilities of k-median method of 3 clusters where the most typical county of each cluster (the cluster center) is shown.



Figures 5: death probabilities for age groups, k-median clustering (k=3)

From age 15 cluster No 3 (Spain, Italy, and France) has the lowest death probabilities; however, in younger ages its top position is not prevalent. The variance in Cluster 2 (East European countries) in younger ages is outstanding.

CONCLUSIONS

In our research we investigated the 2014 death probabilities of 24 European countries and attempted to group them by specific clustering methods. K-means clustering was applied where k=2 gave trivial result, i.e. East and West Europe were differentiated. We increased the cluster number and found that Bulgaria and Estonia are significantly different from other countries of the Eastern block. The Western countries in our research are mainly similar; nevertheless, Spain, Italy, France, and Luxemburg set up a new cluster at k=6 clustering.

We found that k-means method could be problematic while in (re)calculation of cluster centers the *mean* of death probabilities is applied. Since we have significantly different country sizes and also age groups in each country we suggest to use QDEV as an asymmetric similarity measure. K-means method is not suitable to be applied on asymmetric measure; therefore, k-median clustering method was also applied on our dataset. This method gave us exact solution even for higher cluster numbers. K-median clustering drove to the same result as k-means at $k = 2$ clusters but concerning $3 \leq k \leq 6$ results are fairly different. K-median clustering rather divided the group of Western European countries, while k-means separated the Eastern block. We examined former East and West-

Germany separately to identify potential differences left, but we did not find any (even with higher cluster numbers East and West Germany stayed in the same cluster).

For further possible research we might analyze longitudinal changes in death probabilities and find different mortality patterns in the investigated 24 European countries.

ACKNOWLEDGEMENTS

This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00017, titled "Sustainable, intelligent, and inclusive regional and city models".

REFERENCES

Ágoston, K. Cs. (2003). Death rates and their estimation (in Hungarian). In Banyár, J.: Life insurance pp. 377–390. Aula, Budapest.

Ágoston, K. Cs. and Majstorović, S. and Vaskövi, Á. 2019. "Spectral Clustering of Survival Curves". In Proceedings of the 15th International Symposium on Operations Research in Slovenia (ISBN 978-961-6165-55-6), pp. 81-86.

Arató, M., and Bozsó, D. and Elek, P. and Zempléni, A. 2009. "Forecasting and simulating mortality tables." Mathematical and Computer Modelling. Vol. 49, pp. 805-813.

Crocker, K. J., - Snow, A. 2000. The Theory of Risk Classification. In:. Dionne, G. Handbook of Insurance. Kluwer Academic Publishers. Boston / Dordrecht / London.

Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London (Series A), 115:513–585.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [22/01/2020]).

Kovács, E. 2012. "Living Better, Living Longer? Is Ageing in Line with Economic Performance?". Hungarian Statistical Review, Vol. 90, pp. 79-95.

Kovács, E. and Vaskövi, Á. 2019. "Living Longer. Working Longer? Life Expectancy and Retirement Age Trends in OECD Countries". In Proceedings of the 33rd International ECMS Conference on Modelling and Simulation in Italy, pp. 103-108.

Makeham, W. (1867). On the law of mortality. Journal of the Institute of Actuaries, 13(6):325–358.

Marshall, A.W., Olkin, I. 2007. "Life Distributions. Structure of Nonparametric, Semiparametric, and Parametric Families", Springer, New York.

Organization for Economic Cooperation and Development (OECD). 2018. *OECD Pensions Outlook 2018*. OECD Publishing, Paris

Olszewski D. 2011. "Asymmetric k-Means Algorithm". In: Dobnikar A. - Lotrič U. - Šter B. (eds) Adaptive and Natural Computing Algorithms. ICANNGA 2011. Lecture Notes in Computer Science, vol 6594. Springer, Berlin, Heidelberg.

Okada, A. 2000. "An Asymmetric Cluster Analysis Study of Car Switching Data". In: Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Heidelberg.

Vékás. P. 2019. "Rotation of the age pattern of mortality improvements in the European Union". Central European Journal of Operations Research https://doi.org/10.1007/s10100-019-00617-0

AUTHORS' BIOGRAPHIES

KOLOS CS. AGOSTON, graduated as an actuarist and wrote his PhD thesis in insurance markets. He is now an associate professor at Corvinus University of Budapest where he teaches various subjects in operational research and actuarial sciences. He is also the head of Institute of Mathematics and Statistical Modelling. His research topics belong to optimization problems such as cash management, cutting problems and recently college admission problem. His email address is kolos.agoston@uni-corvinus.hu

ÁGNES VASKÖVI, MSc is a PhD candidate at Corvinus University of Budapest, and an assistant professor of the Institute of Finance, Accounting and Business Law. She earned her master's degree in Economics from Corvinus University of Budapest, specializing in financial investment analysis. She gained professional experience in fields of project financing, venture capital and real estate investments. Currently, she teaches Finance, Corporate Finance, and Multivariate Data Analysis. On her main research agenda there are topics of behavioural finance, financial literacy, long term savings, longevity, and pension. Her email address is agnes.vaskovi@uni-corvinus.hu