Central European Journal of Economic Modelling and Econometrics

Principal Component Analysis - Points of Association Between Cancer and Economic Development

Ilona Székely Kovácsné*, Éva Fenyvesi†, Tibor Pintér‡

Submitted: 28.09.2022, Accepted: 15.05.2023

Abstract

Various theories have been put forward on the demographic and health effects and consequences of socioeconomic development. In this study, we used the theoretical findings of the epidemiologic transition as a starting point to examine the 2020 values of the three main cancer indicators (incidence, mortality, prevalence). These values were compared with socioeconomic development variables for 170 countries. The countries were grouped using hierarchical clustering, and linear discriminant analysis was used to evaluate how appropriate the clustering was. Principal component analysis was used to examine, by group, which parameters are significant in each principal component and what background factors underlie the data. The results seem to confirm the association between cancer and socioeconomic background.

Keywords: standard of living, economic development, incidence, mortality, prevalence

JEL Classification: I15, I18, O35, P16

*Budapest Business School, FCHT, Department of Methodology for Business Analysis; e-mail: Kovacsneszekely.Ilona@uni-bge.hu; ORCID: 0009-0000-0025-9125

†Budapest Business School, FCHT, Department of Economics and Business Studies; e-mail: fenyvesi.eva@uni-bge.hu; ORCID: 0000-0002-1500-9409

‡Budapest Business School, FCHT, Department of Economics and Business Studies; e-mail: pinter.tibor@uni-bge.hu; ORCID: 0000-0002-5584-6439

1 Introduction

The process of socioeconomic development and the level of economic development are often interpreted in a one-sidedly positive way in various literature. In health economics, on the other hand, in addition to analyzing investment in health infrastructure, researchers are trying to shed light on the interrelationship between health and economic growth. These studies have increasingly highlighted the fact that certain types of diseases, such as cancer, are much more prevalent in highly developed countries, suggesting that the growth-oriented development model has its downsides (Ukraintseva and Yashin, 2005). The present study is based on the much-criticized theory of epidemiologic transition, which was founded on a 1971 paper by Abdel M. Omran. The theory distinguished three phases of epidemiologic transition based on the mortality patterns of the countries concerned. 1: the era of plague and famine; 2: the era of the declining pandemic; 3: the era of degenerative and man-made diseases. The first era was characterized by fluctuating mortality from epidemics, famine, and war, while the second was marked by endemic non-communicable diseases alongside declining epidemics. In contrast, in the third era, civilizing process and developments in hygiene led, in addition to increasing life expectancy, to the development of civilization-related diseases such as cardiovascular diseases and cancer (Omran, 1971). A study by Olshansky and Ault, published in 1986, adds a fourth era to the previous theory, which the authors call the era of delayed degenerative diseases. It is characterized by a disease pattern resulting from the ageing of Western civilization (Olshansky and Ault, 1986). This group of theories see the phenomenon that epidemics are becoming less and less dominant among the causes of death and that degenerative diseases are becoming more dominant, a consequence of the civilizing process. The theory is supported by cross-sectional, non-time series data: developed countries typically have much higher values for the latter group of patients compared to less developed countries.

The term "health transition" has recently become prominent in the literature (Vallin and Meslé, 2004). The concept of health transition, as a further development of the epidemiologic transition theory, also takes into account how societies react to particular health situations and eras, which socioeconomic periods the different stages of development occurred in, and what events may have amplified the divergence and convergence of mortality in different regions.

Fodor points out that the term "civilization disease" has been used for centuries, and includes various cancers (Fodor, 2013).

The literature reviewed that has identified the socioeconomic background of cancer, presented comparative analysis of international trends based on either secondary (Aggarwal et al., 2014; Arnold et al., 2016; Bos et al., 2005; Ferlay et al., 2018; Teppo, 1984; You et al., 2018) or primary (Donkers et al., 2020; Dumont et al., 2019; Finke et al., 2020; Vehko et al., 2016) data, and predominantly used two approaches. Studies that aggregate different types of cancer using secondary data are emerging as

a distinctive area of research. These studies concentrate on one regional unit (from subnational to supranational level) but may also focus on the comparison between several regional units, with data aggregated to different regional levels. Time is of great importance in cancer research, so most of these studies compare or summarize several years of data. These can be considered the first group of literature (Cutler, 2008; Ferlay et al., 2018; Hofmarcher et al., 2020; Klotz et al., 2019, Luzzati et al., 2018; Steliarova-Foucher et al., 2018).
Other sources (Gunderson et al., 2011; Kanavos and Schurer, 2010; Lawson, 1999; Minicozzi et al., 2018; Ouakrim et al., 2015; Wübker, 2014), on the other hand, present only the socioeconomic background of one or a few types of cancer, typically focusing on differences in incidence and especially mortality using secondary data. This is very important because some types of cancer, such as cervical cancer, have very low rates in higher income countries, while the opposite is true for breast cancer and gastrointestinal cancers.
In this study, our objective is to explore the characteristics of the latest global data on cancer incidence in 170 countries from a sociological, economic, and regional economic perspective.
Further objectives of the study are review some of the important statistics and stochastic relationships of the parameters included in the study; examine the background factors that shape cancer incidence, through measured parameters; group the involved countries of the world; examine differences between groups through parameters, stochastic relationships within groups; examine parameters that influence cancer incidence.

2 Material

The study covered countries surveyed by the Global Cancer Observatory (GCO), using data available on the Cancer Today website (Word Health Organization, 2020). This is the source of the incidence, mortality, and 5-year ASR (Age Standardized Rate) prevalence data for 2020. We chose this year because it is the first year in the database to include cancer indicators for more than 180 countries, whereas previous years only included data for around 60 countries. More recent data are not yet available. All of the involved indicators include all types of cancer. This ASR indicator is a summary measure of the proportion that would have been observed in a given population if it had the same age structure as the reference, thus eliminating the biasing effect of age distribution. Calculated standard cancer incidence and cancer mortality rates are expressed per 100,000 persons, thus allowing for spatial comparisons. Data per 100,000 persons is also the published form of data for prevalence. For all five indicators, the database forms a set of data for the same group of countries. It is important to stress that all GCO indicators are based on estimates. 186 entities (in which some overseas territories of France are indicated as separate countries) are included and the world is the 187^{th} observational unit in the data set. These have

been compared with economic indicators from the International Monetary Fund and the World Bank. However, as the databases of the latter had missing values, 170 observational units remained for the present study.
Incidence is the number of new cancer cases, while mortality is the number of cancer deaths over a given period of time (typically 1 year) in a defined population. Prevalence is the number of individuals in a defined population who have been diagnosed with the disease and are still alive at a given time (i.e., survivors). The definition of all three indicators is from the GCO website. For the latter, 5-year prevalence rates are also calculated, which indicate how many patients are still alive after their diagnosis. Prevalence data were also estimated, but it should be noted that national registers were used as a starting point only for a very limited number of countries. Prevalence data were derived from the incidence rates of the so-called Northern countries between 2006 and 2015 and were adjusted with HDI values.
Based on the literature review of our current study, 10 additional indicators were subjectively selected to be included in our baseline table, but not all of them are for the year 2020. If we look at the time series of these indicators and calculate the memory of the processes (autocorrelation function), we can see that these processes have a long memory, there are no momentous changes within a two-year period. This is because 2020 values are not always available for the various indicators. Three of these indicators were included in the data under review based on the April 2021 edition of the International Monetary Fund's World Economic Outlook (WEO) (International Monetary Fund, 2021) GDP per capita (2020), inflation rate and fiscal balance as a percentage of GDP. From the World Bank (2021) database, total fertility rate for 2018 has been added to the database. The HDI (Human Development Index) values for the year 2019 are from the UNDP (United Nations Development Program) website (Human Development Report, 2021). The latter is a complex indicator of development consisting of four pillars and made up of income, education, and life expectancy data, with a hypothetical set of values between 0 and 1, where a higher value represents a higher level of development. 5 additional indicators are also from the World Bank database for the year 2019: trade-to-GDP ratio, urban population as a percentage of a country's total population, employment in agriculture, employment in industry and employment in services as percentages of total employment.
As these processes are long memory processes, no major changes will occur in a year or two. So there is no "calculation error" if the data are not from one year. Furthermore, this is a multivariate data analysis. 11 parameters were used, so for example in cluster analysis one parameter has little effect on the result.

3 Method

Among the methods used, cluster analysis was applied to show an arrangement of countries where countries with similar characteristics are grouped together (Stockburger, 2016). The clustering was performed using a hierarchical method

with standardized parameters and in accordance with Ward's method (Ward, 1963) with squared Euclidean distance. Ward's method at each step, those two clusters are merged that result in the smallest increase in the overall sum of the squared within-cluster distances as measured by the sum of squares deviation (Norušis, 1993; Romesburg, 1984). Therefore, normality plays no role at this point.
Linear discriminant analysis was used to check how appropriate the clustering was. This analysis provides satisfactory results not only for normally distributed data, but for other types of continuous distributions as well, if the violation is caused by skewness rather than outliers (Kovács et al., 2014). It is a method that looks for the linear combination of data that maximizes intergroup differences while minimizing intragroup differences (Webb, 2002). The linear discriminant functions generated by the method classify countries into groups, and this grouping can be compared with the results of the cluster analysis to obtain the proportion of countries correctly grouped by the linear discriminant functions – it can be used to infer the quality of the clustering (Kovács et al., 2012).
Wilks' λ-statistic provides information on the role of a given parameter in the clustering performed (Afifi et al., 2004). The value of a statistic for a given parameter is calculated as the ratio of the sum of squares within the corresponding group to the total sum of squares. When the Wilks' λ-statistic is 1 for a given parameter, that parameter does not affect the classification. On the other hand, if the value is 0 or close to 0, that parameter has the strongest role in the clustering.
The mathematical goal of principal component analysis (PCA) (Jolliffe, 2002) is to reduce the number of dimensions by taking correlated observed variables and creating uncorrelated probability variables into which the information is compressed. Regarding the question whether the assumption of normal distribution is necessary when applying PCA, the following can be cited: "Principal component analysis can be applied to any distribution of y"(Rencher and Christensen, 2012).
One of the results of applying this method is the correlation of the original parameters with the principal components, which allows us to understand the processes involved in the generation of the data studied.
To determine how well the PCA applies to our data, we can calculate the Kaiser, Meyer, and Olkin measure (KMO) (Kaiser, 1970). The calculations were performed in IBM SPSS Statistics version 25.

4 Results

Among the descriptive statistics (Appendix A.1, Table 1) for the involved 170 countries examined, it is noteworthy that the median - with the exception of the parameter's urban population ratio and services employment ratio - is more or less below the mean.
This is most noticeable for the GDP per capita, 5-year prevalence and agricultural employment ratio parameters. The other significant result is which parameter shows

Table 1: Summary descriptive statistics for countries

	N	Mean	Median	Std. Dev.	Coefficient of variation	Range	Min.	Max.
Incidence	170	183.2	154.8	80.1	0.4	374.0	78.4	452.4
Mortality	170	93.0	88.1	22.3	0.2	124.9	51.3	176.2
5-year prevalence	170	723.6	350.2	774.7	1.1	3116.6	56.0	3172.6
GDP per capita	170	13467.4	4581.6	19322.2	1.4	116667.5	253.6	116921.1
HDI	170	0.7	0.7	0.2	0.2	0.6	0.4	1.0
Fertility r.	170	2.7	2.3	1.3	0.5	5.9	1.0	6.9
Trade to GDP r.	170	63.8	54.3	36.5	0.6	209.0	18.1	227.1
Urban pop. r.	170	59.3	60.2	22.6	0.4	86.8	13.3	100.0
Agric. emp. r.	170	23.7	18.0	21.2	0.9	86.2	0.0	86.2
Industrial emp. r.	170	19.9	19.4	7.9	0.4	51.8	1.9	53.7
Services emp. r.	170	56.4	59.0	17.1	0.3	78.1	10.4	88.5

the highest or lowest variability. GDP per capita has the highest value, followed by 5-year prevalence.

The stochastic relationships are presented with a correlation matrix (Table 2). The strongest correlation is found for the parameters Incidence and 5-year prevalence (0.95). The correlation coefficients for GDP and HDI are significant with Incidence and even higher with 5-year prevalence. The HDI also shows a significantly closer linear relationship with the urban population ratio and services employment ratio parameters. A significant finding is that fertility rate and agricultural employment rate show a stronger (0.76) positive directional relationship with each other, while these two parameters show negative directional relationships with all other parameters examined.

The stochastic relationships are presented with a correlation matrix (Table 2). The strongest correlation is found for the parameters Incidence and 5-year prevalence (0.95). The correlation coefficients for GDP and HDI are significant with Incidence and even higher with 5-year prevalence. The HDI also shows a significantly closer linear relationship with the urban population ratio and services employment ratio parameters. A significant finding is that fertility rate and agricultural employment rate show a stronger (0.76) positive directional relationship with each other, while these two parameters show negative directional relationships with all other parameters examined.

Principal component analysis can be performed on the data, as indicated by the high value of the KMO index (0.727). As expected from the significant correlation coefficients of the parameters, the first principal component explains 56.973% of the variance of the data, while PC2 and PC3 explain only a fraction of this (Table 3).

Table 2: Correlation matrix of the stochastic relationships

	Incidence	Mortality	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.58	0.95	0.70	0.76	-0.61	0.21	0.48	-0.59	0.26	0.62
Mortality		1	0.37	0.06	0.27	-0.28	0.23	0.04	-0.15	0.14	0.13
5-year prevalence			1	0.78	0.79	-0.62	0.21	0.51	-0.62	0.27	0.64
GDP per capita				1	0.71	-0.49	0.14	0.55	-0.57	0.18	0.62
HDI					1	-0.86	0.25	0.70	-0.85	0.57	0.79
Fertility r.						1	-0.31	-0.54	0.76	-0.58	-0.67
Trade to GDP r.							1	0.17	-0.27	0.35	0.17
Urban pop. r.								1	-0.74	0.37	0.74
Agric. emp. r.									1	-0.65	-0.94
Industrial emp. r.										1	0.34
Services emp. r.											1

The stochastic relationships are presented with a correlation matrix (Table 2). The strongest correlation is found for the parameters Incidence and 5-year prevalence (0.95). The correlation coefficients for GDP and HDI are significant with Incidence and even higher with 5-year prevalence. The HDI also shows a significantly closer linear relationship with the urban population ratio and services employment ratio parameters. A significant finding is that fertility rate and agricultural employment rate show a stronger (0.76) positive directional relationship with each other, while these two parameters show negative directional relationships with all other parameters examined.

Principal component analysis can be performed on the data, as indicated by the high value of the KMO index (0.727). As expected from the significant correlation coefficients of the parameters, the first principal component explains 56.973% of the variance of the data, while PC2 and PC3 explain only a fraction of this (Table 3).

Table 3: Result of Principal component analysis

Component	Explained variance (%)	Explained cumulative variance (%)
PC1	56.973	56.973
PC2	12.225	69.198
PC3	11.472	80.670

The correlation of the principal components with the original parameters can be used to understand the background processes involved in generating the measured data. Table 4 shows this so-called component matrix, with correlations between the measured parameters and the first three principal components. The importance of the parameters in each principal component is indicated by the high absolute values (greater than 0.7), which vary between groups.

PC1 is significant for all but three of the parameters examined (mortality, trade to GDP ratio, industrial employment ratio). However, it is important that fertility rate, agricultural employment rate is of opposite sign compared to the others. In PC2 only mortality is significant, while in PC3 the trade to GDP ratio is close to the meaningful level.

Hierarchical cluster analysis using Ward's method (Ward, 1963) and squared Euclidean distance for standardized probability variables was used to group the countries in the study. Based on a dendrogram in our research, it was appropriate to form either three or five groups. The three-group model consists of 27, 50 and 93 countries. This puts 55% of the countries in the third group, masking the differences between the 93 countries. Therefore, it seemed more appropriate to form five groups, in which case the third group could be split into groups of 15, 33 and 45 countries. In the next step, linear discriminant analysis (LDA) was applied. This method can

Table 4: The rates of the component matrix

	Component Matrix		
	PC1	PC2	PC3
Incidence	0.84	0.50	-0.10
Mortality	0.34	0.73	0.41
5-year prevalence	0.85	0.37	-0.19
GDP per capita	0.75	0.11	-0.41
HDI	0.96	-0.06	-0.01
Fertility r.	-0.84	0.10	-0.20
Trade to GDP r.	0.33	0.02	0.68
Urban pop. r.	0.75	-0.31	-0.19
Agric. emp. r.	-0.90	0.32	-0.05
Industrial emp. r.	0.56	-0.39	0.55
Services emp. r.	0.86	-0.22	-0.20

be used to address three objectives. (1) improving how appropriate the clustering is, (2) examining the quality of clustering, (3) graphically illustrating the separation of groups.

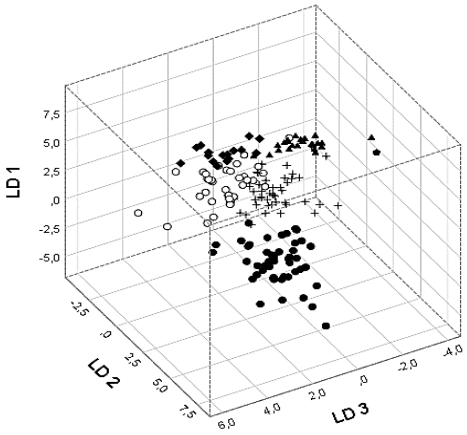
Based on the clustering obtained from the dendrogram (let's call it initial clustering), discriminant functions were defined to test whether each country is in the right group, i.e., to determine how appropriate and reliable a given clustering was. The initial clustering was deemed correct by the discriminant analysis for 93% of countries, suggesting a different, new clustering for 12 countries (7% of cases).

Accepting this, the LDA can be used to examine the extent to which the appropriateness of the clustering has changed in the two cases. With the new clustering proposed by the discriminant analysis, 98.2% of the originally grouped cases were classified correctly. This means that LDA-controlled clustering is so successful that only two countries are not in the correct group according to the discriminant analysis.

The final clustering can be represented by the coordinates of the LD_1, LD_2 and LD_3 functions (Figure 1). The figure shows how the groups are separated. Some overlap slightly. This is because only the first three of the coordinates defined by the parameters are used in the figure to represent each country.

The composition of each clustering (Figure 2) is as follows: 80% of the countries in Group 1 are from Africa, followed by three countries from Asia, two from the Middle East and Oceania and one from the Caribbean. Group 2 is not as homogeneous as the first one. A third of the group is made up of Asian countries, but there are also a significant number of countries from the Middle East (19%) and Central America (15%). The largest part of Group 3 is made up of Asian (18%), South American (25 %) and Caribbean (15%) countries. It includes countries with very large areas, such as Russia, Brazil and China. 75% of Group 4 is made up of EU Member States, joined by the USA, Canada, Australia and New Zealand, as well as Japan and Singapore.

Figure 1: The differentiation of the 5 groups, using coordinates LD1, LD2 and LD3 (one dot indicates a country)



The majority of countries in Group 5 are members of the European Union, but all but one (Portugal) are former socialist countries.
In the correlation matrixes of the groups (see Appendix A.2), there are surprisingly few correlation coefficients that are meaningful (absolute values greater than 0.7). In groups 1 and 2, Incidence shows a strong relationship with Mortality, 5-year prevalence parameters. In groups 3 and 5, Incidence is only strongly related to 5-year prevalence. In groups 3 and 5, 5-year prevalence is correlated with GDP and/or HDI. A further strong negative relationship is found between the Agric. emp. r. and the Services emp. r. (in the first group also the Industrial emp. r.). There are 55 different correlation coefficients in a correlation matrix, 275 in total in the five groups. Only 21 of these can be considered relevant, which is less than 10%. This fact highlights the point that it is more worthwhile to try to use PCA than multivariate regression.
In the correlation matrixes of the groups (see Appendix A.2), there are surprisingly few correlation coefficients that are meaningful (absolute values greater than 0.7). In groups 1 and 2, Incidence shows a strong relationship with Mortality, 5-year prevalence parameters. In groups 3 and 5, Incidence is only strongly related to 5-year prevalence. In groups 3 and 5, 5-year prevalence is correlated with GDP and/or HDI. A further strong negative relationship is found between the Agric. emp. r. and the Services emp. r. (in the first group also the Industrial emp. r.). There are 55 different correlation coefficients in a correlation matrix, 275 in total in the five groups. Only 21 of these can be considered relevant, which is less than 10%. This fact highlights the point that it is more worthwhile to try to use PCA than multivariate regression.

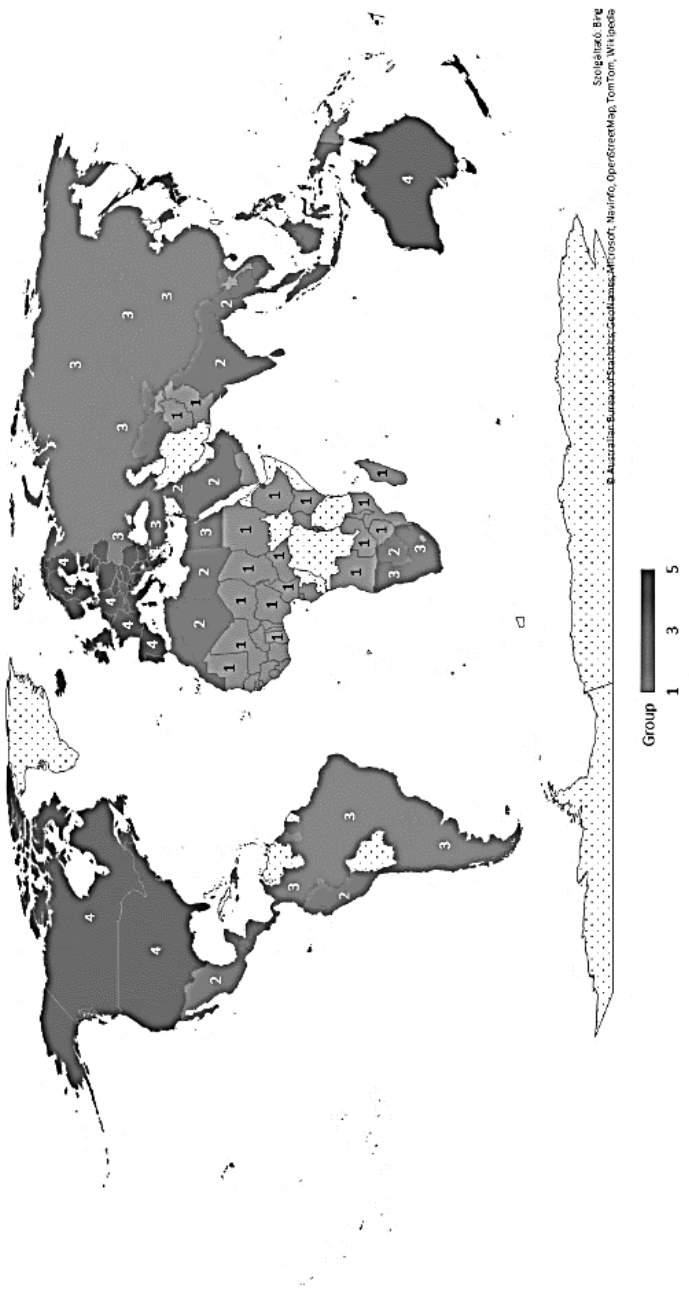


Figure 2: The composition of each clustering

Table 5: The Wilks' λ-statistics of the parameters for the defined grouping

Variable	Wilks' Lambda
5year prevalence	0.106
Incidence	0.158
HDI	0.176
Fertility rate	0.204
GDP per capita	0.291
Rate of agricultural employment	0.352
Rate of services employment	0.401
Rate of industrial employment	0.495
Mortality	0.504
Urban population ratio	0.609
Trade to GDP ratio	0.742

In the correlation matrixes of the groups (see Appendix A.2), there are surprisingly few correlation coefficients that are meaningful (absolute values greater than 0.7). In groups 1 and 2, Incidence shows a strong relationship with Mortality, 5-year prevalence parameters. In groups 3 and 5, Incidence is only strongly related to 5-year prevalence. In groups 3 and 5, 5-year prevalence is correlated with GDP and/or HDI. A further strong negative relationship is found between the Agric. emp. r. and the Services emp. r. (in the first group also the Industrial emp. r.). There are 55 different correlation coefficients in a correlation matrix, 275 in total in the five groups. Only 21 of these can be considered relevant, which is less than 10%. This fact highlights the point that it is more worthwhile to try to use PCA than multivariate regression.

Table 6: KMO values and the variances described by the principal components

Group	KMO Measure of Sampling Adequacy	Extraction Sums of Squared Loadings % of Variance principal component		
		1	2	3
1	0.552	36.069	28.9	9.6
2	0.517	34.100	21.6	13.6
3	0.537	40.233	18.6	9.6
4	0.415	28.061	26.3	12.5
5	0.320	47.029	17.54	14.7

Determining the background factors of the groups provides information on the applicability of principal component analysis. The KMO statistic is under 0.5 for several groups, but the first principal components explain more than 25% of the

variance in the data for all groups. This is also true for the second factor for several groups. The first two factors together explain between 55% and 65% (Table 6).
Table 6 shows the component matrix for the five groups. The first principal components of the first three groups explain between 36 and 40% of the variance of the total data. Group 1 includes the economic sectors considered in the study, in addition to the urban population.

Table 7: The rates of the component matrix

Parameter	Group										
	1		2		3		4			5	
	Component										
	1	2	1	2	1	2	1	2	3	1	2
Incidence	-0.309	0.887	0.038	0.955	0.509	0.737	0.266	0.851	0,167	0.748	0.284
Mortality	-0.460	0.768	-0.356	0.690	-0.394	0.713	0.128	-0.018	0,930	-0.203	0.454
5-year prevalence	0.140	0.956	0.138	0.,879	0.784	0.488	0.133	0.884	0,053	0.934	0.022
GDP per capita	0.601	0.360	0.802	-0.291	0.788	0.031	0.690	0.118	-0,326	0.901	-0.003
HDI	0.663	0.514	0.848	0.234	0.834	0.268	0.576	0.449	-0,016	0.922	0.044
Fertility r.	-0.375	-0.642	-0.186	-0.191	-0.592	-0.169	0.264	0.800	0,122	0.549	0.167
Trade to GDP r.	0.329	0.181	0.001	-0.089	-0.352	0.345	0.418	-0.448	0,546	0.474	0.645
Urban pop. r.	0.701	-0.081	0761	0.061	0.572	-0.135	0.542	-0.382	-0,004	0.420	-0.418
Agric. emp. r.	-0.917	0.131	-0.891	-0.037	-0.808	0.271	-0.619	0.037	0,240	-0.879	0.052
Industrial emp. r.	0.728	-0.125	0.596	-0.128	-0.128	0.503	-0.666	0.322	-0,051	0.071	0.844
Services emp. r.	0.853	-0.113	0.691	0.121	0.780	-0.439	0.900	-0.291	-0,090	0.747	-0.554

It is noteworthy that the role of agriculture is the opposite of the other parameters. In Groups 2 and 3, the important parameters in the first principal components are, similarly to Group 1, some important parameters of the economy, but in these groups their importance decreased, while the role of parameters indicating the well-being of society, such as DGP and HDI, is significant. For Group 4, only the role of services is highly significant. In Group 5, in addition to economic and welfare parameters, cancer-specific parameters also appear in the first factor. It should be noted that the 5-year prevalence also appears in the first principal component of Group 3. In Group 1, only cancer-related parameters are characteristic in the second factor. In Groups 2, 3 and 4, it varies which morbidity-related parameters remain significant. However, the role of mortality varies most strongly from group to group. This goes so far as to include mortality in the third factor as an independent and significant parameter in Group 4 (Table 7).
The box-whiskers plots (Appendix A.3) of the examined parameters for the countries in the groups show a similar pattern from group to group for incidence, 5-year prevalence, HDI, GDP, urban population, and the percentage of employment in services. These plots show that the values of these parameters are lowest in the first group, increase with the number of groups (as a denomination) until Group 4,

and then decrease in Group 5. This means that incidence and 5-year prevalence are similar across groups, as are HDI, GDP, urban population, and the percentage of employment in services. Thus, the smallest values are in Group 1 and the largest values are in Group 4. A virtually opposite pattern is found for the role of agriculture in the economy, with the largest role in Group 1 and the smallest in Group 4 countries. In Group 5, the role of agriculture is increased compared to Group 4 countries and in many cases, reaches the importance of Group 3 countries. The variation in fertility rate values from group to group is similar to this pattern. The values of the groups for the share of industry and trade of goods vary differently from the previous patterns, but similarly from group to group. They are lowest in Group 1 and highest in Group 5. The trend in mortality values from group to group is not similar to any of the parameters.

5 Summary and discussion

The economic context of incidence, mortality and prevalence has been the subject of much research. The present study fills a gap by breaking down almost every country in the world into different groups using cluster analysis, and principal component analysis has allowed these groups to be examined separately through their specific parameters (Table 8). Thus, in the present research, the focus was not on quantitative relationships and changes over time, but on the parameters that describe the different groups and whether these parameters act in the same direction or in opposite directions.

Wilks' λ-statistics show the role of a parameter in clustering (Table 5). 5-year prevalence and incidence both have low values, and thus have the most significant clustering power among the probability variables. This is because these two parameters show the largest differences between groups of countries. The values of Wilks' λ-statistics for HDI and GDP give an indication of the significant differences in living standards across the world. The lower fertility rate highlights how childbearing varies across countries and regions. The also not high Wilks' λ-statistic of agriculture highlights its role in the economy of specific groups and the differences in the world. Of note is the high Wilks' λ value for mortality, i.e., its relatively low clustering significance, especially as 5-year prevalence and incidence have low values and thus have the largest clustering role.

The parameters with the highest values – i.e., the ones that influenced clustering the least – are urban population and trade-to-GDP ratio. Both point to global trends, one of which is the increasing proportion of the world's population living in large cities. The other one is the growing proportion of trade in the economy, no matter how developed the given economy is. A Hungarian study also confirms that world trade has changed radically in the last 50 years, with a more than thousand-fold increase in volume. Various bi- and multilateral preferential trade agreements, technological

innovations, lower transport costs and new types of production processes have all contributed to the ability of all countries to participate in world trade (Vakhal, 2016). The first column of the component matrix of the first group, showing which parameters are important in the first principal component.

However, agriculture, similarly to the other groups and reflecting the world trend, is undergoing the opposite evolution as the economy as a whole: its share of the economy is declining, as indicated by the negative value of the correlation coefficient. Group 2 consists of Asian, Middle Eastern and Central American countries where the role of industry and services is declining but the role of living standards is increasing (GDP and HDI are becoming significant). In the first two groups, the share of urban population has a significant role. It is no longer significant in the other groups, indicating that the growth of the share of urban population in these groups has slowed down. In the first principal component of Group 3, 5-year prevalence of cancer appears in addition to parameters expressing well-being, and the proportion of employment of services and agriculture. This phenomenon is even more pronounced in Group 5, which is the group of former socialist countries, because incidence also appears alongside the previously mentioned parameters. However, it is important to see that disease-specific parameters "usually" appear in the second principal component. In the first two groups, the second factor is disease. In Group 3, incidence and mortality together are the most significant in the second factor. In Group 4, mortality is not involved at all in the second factor but appears instead in the third factor. Based on these results, it can be said that in less developed (Group 1) and developing or more developed (Groups 2 and 3) countries, as the economy develops, tumor incidence also appears in conjunction with mortality. In the most developed countries, mortality is detected in another background factor. In other words, the first principal components express the state of economic development, while the second principal components express disease. However, in the most developed countries, resources are already available and so many resources are being directed toward early detection and treatment that not everyone dies shortly after being diagnosed.

Group 5 should be interpreted carefully, because it includes 15 countries, and 11 parameters are considered. It is unfortunate that the number of parameters does not significantly exceed the number of countries. Taking this into account, it can be concluded that the parameters describing disease incidence are associated with certain parameters of well-being and economy – a situation similar to that of "medium developed countries" – but mortality is no longer associated with disease-specific parameters. In other words, Groups 4 and 5 have the right financial situation and attitude, as well as infrastructure to detect the disease early and "save" (some of) the patients.

Most of the countries within each group formed by the cluster analysis fall within the interquartile range for most parameters.

Table 8 shows the order of the average values of the parameters used. Although the

Table 8: The order of the averages calculated from the values of the different parameters in each country group

Indicators \Groups	1.	2.	3.	4.	5.
Incidence	1	2	3	5	4
Mortality	2	1	4	3	5
5-year prevalence	1	2	3	5	4
GDP per capita	1	2	3	5	4
HDI	1	2	3	5	4
Fertility rate	5	4	3	1	2
Trade to GDP ratio	1	4	2	3	5
Urban population ratio	1	2	4	5	3
Agric. emp. r.	5	4	3	1	2
Industrial emp. r.	1	4	3	2	5
Services emp. r.	1	2	4	5	3

ordering blurs the magnitude of the real values, it facilitates the comparison of the groups.

Table 8 shows that, in terms of economic parameters, countries in Group 1 are the least developed and those in Group 4 are the most developed. Group 5 includes Portugal and former socialist countries. Although the economic structure of the former socialist countries has changed a lot since the change of regime, the specific development path has left its mark to this day. For example, the share of industry is still markedly significant within the economy. In these countries, mortality rates are particularly high, but higher GDP per capita means that they can spend more on healthcare, so prevalence rates are also high. Group 3 has the most varied findings. Although lagging Group 5 in terms of GDP per capita and HDI, the share of industry and services, and their large urban population show signs of development. In their case, caution must be exercised when it comes to GDP per capita, as this group includes many countries with large populations, such as China, Brazil, Egypt, and Russia. Regarding the composition of the groups formed during the cluster analysis, Groups 4 and 5 are made up of mostly Western European and North American, and Central and Eastern European countries, respectively, with a distinct separation. This confirms the similar classification of previous works also studying groups of countries (Egri, 2017b; Ferlay et al., 2018; Steliarova-Foucher et al., 2018; Minicozzi et al., 2018; Arnold et al., 2016; Hofmarcher et al., 2020; Ouakrim et al., 2015).

Our research shows that incidence has a predominant role in high-income, developed European and North American countries, Australia and New Zealand (Group 4). This is confirmed by several studies. These studies have shown a high incidence in developed countries (Teppo, 1984; Ukraintseva and Yashin, 2005; Ferlay et al., 2018; Luzzati et al., 2018; Minicozzi et al., 2018).

Although we examined the relationship between parameters at one point in time, but incidence is an important parameter for the main components of the groups

– and our other knowledge of the groups (e.g., developed, underdeveloped, former socialist countries) suggests that economic development plays a significant role in cancer incidence.

This is also indicated by various time series studies showing that the number of disease cases is steadily increasing as the economy and living standards improve. Arnold and colleagues (2015) examined cancer incidence between 1988 and 2008 for 26 European countries. They found that the initially lower values in the Central and Eastern European region have become higher and higher, catching up with higher-income European countries. In another study based on the World Health Organization's (WHO) mortality database for forty-three countries, Arnold et al. found that in more developed countries, incidence is higher for all types of cancer except cervical cancer (development expressed as HDI – Human Development Index) (Arnold et al., 2016). Luzzati et al. expressed the level of development as GDP per capita and found a positive relationship between incidence and income in 122 countries, which remained positive even after adjusting for control variables (Luzzati et al., 2018). The link between higher incidence and higher living standards is also demonstrated by the research of Amin and Rivera (2020). Spatial clusters of incidence and mortality were formed for oral and pharyngeal cancer in the states of the United States of America. In terms of incidence, coastal metropolitan agglomerations showed a higher incidence. In 2014, Aggarwal et al. examined health economic trends in cancer incidence in the EU countries, Canada, and the United States. The authors of the study used two-dimensional plots to show that incidence has a slightly positive correlation with income, while the correlation with mortality is strongly negative. They also found that mortality data are the worst in middle and upper middle-income countries rather than in countries with the highest incidence. They also found that over the time period studied (1975–2010), mortality rate has steadily decreased, and incidence rate has increased, thus increasing the gap between the two priority indicators (Aggarwal et al., 2014). Although in our study, parameter expressing well-being are not among the background parameters in the group of developed countries (Group 4), it can be concluded that the results of the two studies are similar: in our results, mortality in developed countries is not associated anymore with incidence and 5-year prevalence. Another study (Bos et al., 2005) evaluated cancer incidence data from the Netherlands and found that cancer mortality is essentially low and thus survival is high in regions where socioeconomic and income inequality is low. This study draws attention to the inclusion of another parameter in further studies, as our research does not include directly social and income inequality.

A 2018 study also aimed to identify patterns of incidence and mortality based on European countries (Ferlay et al., 2018). The starting point of the study was that the European continent accounts for 9% of the world's population, but 25% of all cancer patients. The researchers have established indicators for 40 European countries and put these countries into four major regional groups: Central and Eastern Europe, Northern Europe, Southern Europe, Western Europe. In our study, we followed a

different methodological path. Our groups were created mathematically based on the parameters describing the countries. In a study by Ferlay and colleagues (2018), Hungary was found to have the highest incidence of cancer in all countries. In addition to Hungary, Estonia, France, Ireland, Latvia, Norway, Slovakia, and Slovenia also had high incidence rates. The countries with the lowest incidence rates were Albania, Bosnia and Herzegovina, Ukraine. The countries with the highest cancer mortality rates included Hungary, Slovakia, Latvia, Lithuania, Serbia, and Croatia. However, Sweden, Finland, Albania, and Spain had the lowest mortality rates. It is important to note that the countries in the groups we have formed do not match the countries in the groups in the study by Ferlay et al. (2018). Norway and France are in a different group according to our results, but it occurred due to different parameters and our data are about a later period, 2020. This is probably because we are not looking with PCA at mortality or prevalence, but at how the parameters included in the study are related. However, in Group 5, incidence, 5-year prevalence, GDP and HDI are all important parts of the first principal component. Our results show that mortality diverges from incidence as living standards rise (mortality is included in a separate principal component, explaining only 12.5% of the data). This is particularly true in the most developed countries, where mortality has already completely diverged from incidence. The reason behind this is that they have enough money to spend on prevention (research, screening, care, etc.). For the Central and Eastern European region (Group 5), incidence, 5-year prevalence, GDP per capita and HDI are in the first principal component, explaining 47% of the variance in this group of data. This suggests that disease is numerically significant. It should be noted that mortality is much more significant in this group than in Group 4, which includes developed European countries. In other words, incidence and prevalence are increasing, but now, the latter is much less effective in the former socialist countries than in the western part of Europe, and therefore mortality is high.

Many studies do not examine mortality in relation to cancer as a whole, but in relation to specific diseases. Minicozzi et al. investigated the incidence, mortality and survival of pancreatic cancer and bile duct cancer (Minicozzi et al., 2018). Data from 29 European countries were also analyzed for the period between 2000 and 2007, broken down by sex, age group and region (UK and Ireland; Northern; Central; Southern, Eastern Europe). The worst situation was found in Eastern Europe, especially in terms of mortality, but survival also declined with age. However, truly significant difference compared to the other regions was not found, a trend of increasing incidence was present everywhere. Kanavos and Schurer's study examined surveillance of colorectal cancer in 17 countries, a process that is considered important mainly for earlier diagnosis (Kanavos and Schurer, 2010). A study 5 years later (Ouakrim et al., 2015) has examined colorectal cancer mortality in as many as 34 European countries. It concluded that significant reduction was seen in countries where the availability of surveillance, screening and specialized treatment is good (e.g., USA). They also showed that women have better survival prognosis for this type of disease

compared to men. A study in Australia found that children of higher socioeconomic status, who were taller than poorer children and had a higher average body weight, had a 10% higher breast cancer mortality rate after three and a half decades, and incidence rates were even higher (Lawson, 1999). A paper published in 2011 suggests that the incidence of prostate cancer is higher in areas where Northern European Viking populations migrated to and established new settlements, suggesting a genetic background (Gunderson et al., 2011). The authors also point to a trend difference between the raw incidence rate and the ASR (age standardized rate) with territorial and ethnic bases in their perspective. Another study focused on the incidence of testicular cancer in a 35-year time interval and included data from 41 countries (Gurney et al., 2019). According to their research, the incidence of the disease is very high in Western European and Northern European countries, and lower in less developed countries, but it is on the rise there, as well.

Klotz and colleagues have looked at projections of incidence and mortality rates up to 2030 (Klotz et al., 2019). The authors of the study expect a clear increase in incidence in Austria, with lifestyle factors playing a major role in addition to ageing. Using a quasi-Poisson regression model, they mainly incorporated the effects of ageing into their estimates, but also pointed out that regional differences at the provincial level should be taken into account. Our results place Austria in Group 4, with incidence as an important parameter in the group's second principal component.

However, the results of a study in one of the less developed socialist countries (Group 5) (a health inequality study in Hungary) suggest that health in higher-income metropolitan areas is more favorable (Egri, 2017a). However, the link between cancer and development also depends on the type of cancer discussed. For example, cervical cancer is a type of cancer with an overall poorer outlook in lower-income countries (Arbyn et al., 2020), in which the spread of HPV vaccination in more developed countries has probably played a major role (Ferko et al., 2008).

For three of the groups we examined, it can be stated that there is a smaller or stronger relationship between fertility rates and incidence. However, this relationship is far from parallel. For Group 1, the incidence rate is 0.88 and the fertility rate is -0.64. In Group 4, both parameters play a significant role in the second principal component and have the same sign, i.e., larger families, higher incidence. In Group 5, incidence in the first principal component is present with the same sign as fertility rate, similarly to Group 4. However, in Groups 2 and 3, there is no correlation between these two parameters. These data also confirm that it is especially important to pay attention to which groups of countries are being examined when looking at the relationships between these parameters. For example, the link between total fertility rate and cancer incidence data in 178 countries examined by You and colleagues (2018) presents average results and does not allow us to detect differences between countries/country groups. GDP per capita, life expectancy and the percentage of urban population were used as control variables. Their results show a moderately strong negative correlation between family size and incidence. It was concluded that

the emotional balance of communities with larger average family sizes creates an environment that can provide individuals with higher resistance to cancer. In our case, this is only true for countries in Groups 4 and 5.

In addition to the direct health costs associated with cancer, they have quantified the fertility losses due to the increasing number of people affected with the disease in 31 European countries. Their calculations show that Germany, Denmark, Switzerland, the Netherlands, Belgium and Luxembourg have the highest per capita costs (per capita costs expressed in euros, adjusted for purchasing power parity) (Hofmarcher et al., 2020). This confirms our finding that Group 4, which includes developed countries, has a very low fertility rate along with high incidence and prevalence. The close relationship between these parameters is evidenced by the fact that all of them play a significant role in the second principal component for this group.

As mentioned earlier, a limitation in comparing our results with other studies is that we did not use a time series and omitted some demographic characteristics such as age and sex. By including these parameters, research on cancer can be further refined. For example, a study by Cutler (2008) also highlights trends in cancer-specific mortality, specifically that mortality in the US increased steadily from 1933 to 1993, but then began to decline steadily, mainly due to improved diagnostic and treatment methods (Cutler, 2008). However, the ability of patients to pay for their treatment also plays an important role.

More complex statistical studies in the future could also provide answers to the question of what environmental, social, lifestyle and genetic characteristics (Gunderson et al., 2011, Amin and Rivera, 2020) are responsible for the fact that not all high-income countries have similar incidence and prevalence rates. In the context of mortality, our results differ from the results presented earlier (Arnold et al., 2016; Bos et al., 2005) in that there is no clear relationship with socioeconomic indicators, thus mortality is not lower in higher income countries globally (Munro, 2014), but it is not higher either as the clustering suggests. Indeed, our research shows that for countries in Groups 4 and 5, mortality diverges from incidence and five-year prevalence. Group 5, however, has a much higher mortality rate than Group 4 countries. Countries in Group 4 are less developed than those in Group 5, but the latter are ranked second among the clustered groups in terms of GDP per capita. Mortality is therefore not only a question of GDP per capita, but also of healthcare quality, healthcare equipment etc. in each country... The former socialist countries (Group 5) still fall behind Group 4 in several areas.

During the interpretation of certain study results, it is also important to consider whether those studies have narrowed their dataset to a specific country (Ferlay et al., 2018; Steliarova-Foucher et al., 2018), because in this case there may be a stronger statistical relationship between the different parameters than in our work, where we included data from all possible countries.

Regarding further studies, we also consider it important to investigate whether there are determinants of cancer related to socioeconomic development, the identification

of which could lead to changes in economic policy and development models. It can highlight that in countries with higher living standards, there is a demographic shift towards ageing and urbanization. Ageing societies together with higher life expectancy (Micheli et al., 2003; Molnár and Barna, 2012) and urbanization trends, favor the development of cancer (Di et al., 2015).
The economic development process, as we understand it today, can therefore also function as a significant risk factor for the development and spread of cancer, as has been published previously (Ukraintseva and Yashin, 2005). However, the decrease in prevalence is not clearly a positive trend, as it may also be a consequence of higher mortality, while higher incidence rates may be due to better detection, screening, and registration techniques, and do not necessarily reflect higher case incidence.
The values appearing on the mortality side also appear earlier on the incidence side, which is why, in our opinion, incidence data could not be distorted to such an extent that it cannot be said that *cancer is a civilization disease, and its occurrence is related to socioeconomic development*. Statistically, incidence and five-year prevalence show the clearest positive co-movement with living standards (Table 5). Based on the PCA, baseline and 5-year prevalence are in the same principal component with the same sign (except for Group 3). The main reason for this is that even in developed countries, cancer treatment is not very effective today.

6 Conclusions

The study used data from the first year of the GLOBOCAN database, which is available for a significant proportion of the world's countries, this is the year 2020. As the result of our research, we have been able to identify 5 groups of countries, each with different characteristics. The groups were created mathematically based on the variables included in the study. There are few high-income countries in the world where the incidence or prevalence rates are not high. This suggests that cancer can be considered a disease of civilization. Cancer indicators are essentially based on estimates, so our conclusions are of course open to debate. In the following years, we can expect to create more data for more than 170 countries of the world - this will also facilitate result comparisons over time.

References

[1] Afifi A., Clark V. A., May S., Raton B., (2004), *Computer-Aided Multivariate Analysis (4th ed.)*, Chapman & Hall/CRC, 489 p. ISBN 1-58488-308-1, USA.

[2] Aggarwal A., Ginsburg O., Fojo T., (2014), Cancer economics, policy and politics: What informs the debate? Perspectives from the EU, Canada and US, *Journal of Cancer Policy* 2, 1–11.

[3] Amin R. W., Rivera B., (2020), A spatial study of oral and pharynx cancer mortality and incidence in the U.S.A.: 2000-2015, *Science of the Total Environment* 713, 136688.

[4] Arbyn M., Weiderpass E., Bruni L. et al., (2020), Estimates on incidence and mortality of cervical cancer in 2018: a worldwide analysis, *Lancet Global Health* 8, 191–203.

[5] Arnold M., Rentería E., Conway D. I. et al., (2016), Inequalities in cancer incidence and mortality across medium to highly developed countries in the twenty-first century, *Cancer Causes Control* 27, 999–1007

[6] Bos V., Kunst A. E., Garsse, J., Mackenbach J. P., (2005), Socioeconomic Inequalities in Mortality within Ethnic Groups in the Netherlands, 1995-2000, *Journal of Epidemiology and Community Health* (1979-) 59(4), 329–335.

[7] Cutler D. M., (2008), Are We Finally Winning the War on Cancer?, *The Journal of Economic Perspectives* 22(4), 3–26.

[8] Di J., Rutherford S., Chu C., (2015), Review of the Cervical Cancer Burden and Population-Based Cervical Cancer Screening in China, *Asian Pacific Journal of Cancer Prevention* 16, 7401–7407.

[9] Donkers H., Bekkers R., Massuger L., Galaal K., (2020), Socioeconomic deprivation and survival in endometrial cancer: The effect of BMI, *Gynealogic Oncology* 156, 178–184.

[10] Dumont S., Cullati S., Manor O. et al., (2019), Skin cancer screening in Switzerland: Cross-sectional trends (1997–2012) in socioeconomic inequalities, *Preventive Medicine* 129, 105829.

[11] Egri Z., (2017a), Magyarország városai közötti egészségegyenlőtlenségek, *Területi Statisztika* 57(5), 537–575.

[12] Egri Z., (2017b), Térségi egészségegyenlőtlenségek az európai makrorégióban (kelet-közép-európai szemszögből), *Területi Statisztika* 57(1), 94–124.

[13] Ferko N., Postma M., Gallivan S. et al., (2008), Evolution of the health economics of cervical cancer vaccination, *Vaccine* 26S, F3–F15.

[14] Ferlay J., Colombet M., Soerjomataram I. et al., (2018), Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018, *European Journal of Cancer* 103, 356–387.

[15] Finke I., Behrens G., Schwettmann L. et al., (2020), Socioeconomic differences and lung cancer survival in Germany: Investigation based on population-based clinical cancer registration, *Lung Cancer* 142, 1–8.

[16] Fodor L., (2013), A civilizációs betegségek pszicholó1giai körvonalai, available at: http://rmpsz.ro/uploaded/tiny/files/magiszter/2013/tel/3.pdf, access 16.05.2021.

[17] Gunderson K., Wang C. Y., Wang R., (2011), Global prostate cancer incidence and the migration, settlement, and admixture history of the Northern Europeans, *Cancer Epidemiology* 35, 320–327.

[18] Gurney J. K., Florio A. A., Znaor A. et al., (2019), International Trends in the Incidence os Testicular Cancer: Lessons from 35 years and 41 Countries, *European Urology* 76, 615–623.

[19] Hofmarcher T., Lindgren P., Wilking N., Jönsson B., (2020), The cost of cancer in Europe 2018, *European Journal of Cancer* 129, 41–49.

[20] Human Development Report, (2021), available at: http://hdr.undp.org/en/data, access 20.05.2021.

[21] International Monetary Fund, (2021), available at: https://www.imf.org/en/Publications/WEO/weo-database/2021/April, access 16.05.2021.

[22] Jolliffe I. T., (2002), *Principal Component Analysis*, Springer Series in Statistics, Second Edition, 405 p.

[23] Kaiser H. F., (1970), A second generation little jiffy, *Psychometrika* 35(4), 401–415.

[24] Kanavos P., Schurer W., (2010), The dynamics of colorectal cancer management in 17 coutries, *The European Journal of Health Economics* 10(1), 115–129.

[25] Klotz J., Hackl M., Schwab M., Hanika A., Haluza D., (2019), Combining population projections with quasi-likelihood models, *Demographic Research* 40, 503–532.

[26] Kovács J., Nagy M., Czauner B., Kovács I. Sz., Borsodi A. K., Hatvani. I. G., (2012), Delimiting sub-areas in water bodies using multivariate data analysis on the example of Lake Balaton (W Hungary), *Journal of Environmental Management* 110, 151–158.

[27] Kovács J., Kovács S., Magyar N., Tanos P., Hatvani I. G., Anda A., (2014), Classification into homogeneous groups using combined cluster and discriminant analysis, *Environmental Modelling and Software* 57, 52–59.

[28] Lawson J. S., (1999), The link between socioeconomic status and breast cancer-a possible explanation, *Scandinavian Journal of Public Health* 27(3), 203–205.

[29] Luzzati T., Parenti A., Rughi T., (2018), Economic Growth and Cancer Incidence, *Ecological Economics* 146, 381–396.

[30] Micheli A., Baili P., Quinn M., Mugno E., Capocaccia R., Grosclaude P., (2003), Life expectancy and cancer survival in the EUROCARE-3 cancer registry areas, *Annals of Oncology* 14, 28–40.

[31] Minicozzi P., Cassetti T., Vener C., Sant M., (2018), Analysis of incidence, mortality, and survival for pancreatic and biliary tract cancers across Europe, with assessment of influence of revised European age standardisation on estimates, *Cancer Epidemiology* 55, 52–60.

[32] Molnár T. M., Barna K., (2012), Demográfiai jellemzők Magyarországon és az Európai Unióban, különös tekintettel a daganatos megbetegedések okozta halálozásra, *Statisztikai Szemle* 90(6), 544–558.

[33] Munro A. J., (2014), Comparative cancer survival in European countries. *British Medical Bulletin* 110, 5–22.

[34] Norušis M. J., (1993), *SPSS for Windows Base System Users Guide*, Release 6.0. SPSS Inc., Chicago.

[35] Olshansky S. J., Ault A. B., (1986), The Fourth Stage of the Epidemiologic Transition: The Age of Delayed Degenerative Diseases, *The Milbank Quarterly* 64(3).

[36] Omran Abdel E., (1971), The Epidemiologic Transition: A Theory of the Epidemiology of Population Change, *The Milbank Memorial Fund Quarterly* 49(4), 509–538.

[37] Ouakrim D. A., Pizot C., Boniol M. et al., (2015), Trends in colorectal cancer mortality in Europe: retrospective analysis of the WHO mortality database, *British Medical Journal* 351.

[38] Rencher A. C., Christensen W. F., (2012), *Methods of Multivariate Analysis*, 3rd Edition,Wiley Series in Probability and Statistics, Singapore.

[39] Romesburg H. C., (1984), *Cluster Analysis for Researchers*, Belmont, Calif.: Lifetime Learning Publications.

[40] Steliarova-Foucher E., Fidler M. M., Colomet M. et al., (2018), Changing geographical patterns and trends in cancer incidence in children and adolescents in Europe, 1991–2010 (Automated Childhood Cancer Information System): a population-based study, *Lancet Oncology* 19, 1159–69.

[41] Stockburger D. W., (2016), *Introductory Statistics: Concepts, Models, and Applications*, Missouri State University, 3rd Web Edition.

[42] Teppo L., (1984), Cancer incidence by living area, social class and occupation, *Scandinavian Journal of Work, Environment & Health* 10(6), 361–366.

[43] The World Bank, (2021), available at: https://data.worldbank.org/ access 18.05.2021.

[44] Ukraintseva S. V., Yashin A. I., (2005), Economic Progress as Cancer Risk Factor II. Why is Overall Cancer Risk Higher in More Developed Countries? MPIDR Working Paper, Rostock.

[45] Vakhal P., (2016), A hozzáadott-érték kereskedelem tendenciái az OECD országokban. Budapest: KOPINT-TÁRKI Konjunktúrakutatási Intézet Zrt, available at: https://www.parlament.hu/documents/126660/712568/TiVA_v2_a.pdf

[46] Vallin J., Meslé F., (2004), Convergences and divergences in mortality. A new approach to health transition, *Demographic Research, Special Collection* 2, 11–44.

[47] Vehko T., Arffman M., Manderbacka K. et al., (2016), Differences in mortality among women with breast cancer by income – a register-based study in Finland, *Scandinavian Journal of Public Health* 44(7), 630–637.

[48] Ward J. H., (1963), Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58(301), 236–244.

[49] Webb A. R., (2002), *Statistical Pattern Recognition*, Second Edition, John Wiley & Sons.

[50] World Health Organization, (2020), available at: https://gco.iarc.fr/today/home, access 16.05.2021.

[51] Wübker A., (2014), Explaining variations in breast cancer screening across European countries, *The European Journal of Health Economics* 15(5), 497–514.

[52] You W., Rühli F. J., Henneberg R. J., Henneberg M., (2018), Greater family size is associated with less cancer risk: an ecological analysis of 178 countries, *BMC Cancer* 18, 924.

Appendix A

A.1 Descriptive statistics by group

Table A1: Descriptive statistics in group 1

Group 1	N	Mean	Median	Std. Deviation	Coefficient of variation	Range	Minimum	Maximum
Incidence	50	119.42	114.85	26.75	0.22	122.0	78.4	200.4
Mortality	50	84.90	81.30	18.22	0.21	84.6	54.8	139.4
5-year prevalence	50	128.46	121.15	42.32	0.33	192.5	56.0	248.5
GDP per capita	50	1586.48	1151.03	1419.64	0.89	7167.6	253.6	7421.2
HDI	50	0.54	0.54	0.07	0.14	0.3	0.4	0.7
Fertility rate	50	4.34	4.41	0.97	0.22	5.0	1.9	6.9
Trade to GDP rate	50	46.52	44.46	15.26	0.33	65.8	18.1	83.9
Urban population ratio	50	40.47	37.57	17.47	0.43	76.5	13.3	89.7
Rate of agricultural employment	50	49.13	49.24	17.00	0.35	67.9	18.3	86.2
Rate of industrial employment	50	12.19	11.37	5.53	0.45	23.3	1.9	25.2
Rate of services employment	50	38.68	39.47	13.42	0.35	53.9	10.4	64.3

Table A2: Descriptive statistics in group 2

Group 2	N	Mean	Median	Std. Deviation	Coefficient of variation	Range	Minimum	Maximum
Incidence	45	131.44	133.50	21.56	0.16	85.3	91.0	176.3
Mortality	45	76.50	75.30	12.27	0.16	54.7	51.3	106.0
5-year prevalence	45	305.47	255.10	132.94	0.44	509.3	134.4	643.7
GDP per capita	45	7671.83	3988.45	9361.93	1.22	51141.2	1003.0	52144.2
HDI	45	0.73	0.74	0.08	0.12	0.4	0.5	0.9
Fertility rate	45	2.34	2.31	0.48	0.20	2.1	1.4	3.5
Trade to GDP rate	45	72.58	64.32	41.97	0.58	200.6	26.5	227.1
Urban population ratio	45	60.06	62.99	22.68	0.38	81.4	18.6	100.0
Rate of agricultural employment	45	19.89	17.37	12.66	0.64	43.4	0.9	44.3
Rate of industrial employment	45	24.09	23.59	7.56	0.31	40.4	13.3	53.7
Rate of services employment	45	56.02	57.43	11.64	0.21	43.8	32.3	76.1

Table A3: Descriptive statistics in group 3

Group 3	N	Mean	Median	Std. Deviation	Coefficient of variation	Range	Minimum	Maximum
Incidence	33	203.79	198.80	30.32	0.15	130.8	159.4	290.2
Mortality	33	110.53	107.60	19.55	0.18	96.1	80.1	176.2
5-year prevalence	33	664.98	642.10	279.93	0.42	1108.0	271.8	1379.8
GDP per capita	33	9433.81	6783.05	8770.16	0.93	41997.9	1690.7	43688.6
HDI	33	0.77	0.78	0.06	0.08	0.3	0.6	0.9
Fertility rate	33	2.17	1.98	0.63	0.29	2.6	1.3	3.9
Trade to GDP rate	33	52.12	49.48	21.89	0.42	80.0	22.3	102.3
Urban population ratio	33	66.26	66.86	18.43	0.28	77.4	18.1	95.4
Rate of agricultural employment	33	14.23	13.82	9.76	0.69	38.1	0.1	38.2
Rate of industrial employment	33	20.73	20.76	3.84	0.19	13.3	14.1	27.4
Rate of services employment	33	65.04	67.60	10.79	0.17	38.4	44.9	83.3

Table A4: Descriptive statistics in group 4

Group 4	N	Mean	Median	Std. Deviation	Coefficient of variation	Range	Minimum	Maximum
Incidence	27	309,60	292,60	54,99	0,18	219,40	233,00	452,40
Mortality	27	94,75	93,50	10,78	0,11	38,20	75,50	113,70
5-year prevalence	27	2128,89	2158,80	515,99	0,24	1988,60	1184,00	3172,60
GDP per capita	27	49325,85	45732,80	21684,84	0,44	99250,83	17670,29	116921,11
HDI	27	0,92	0,93	0,02	0,03	0,09	0,86	0,96
Fertility r.	27	1,51	1,52	0,22	0,15	0,90	0,98	1,88
Trade to GDP r.	27	66,76	54,65	41,00	0,61	181,94	19,65	201,59
Urban pop. r.	27	82,37	82,62	10,55	0,13	41,49	58,52	100,00
Agric. emp. r.	27	3,01	2,53	2,31	0,77	11,57	0,03	11,60
Industrial emp. r.	27	19,96	19,29	3,63	0,18	16,37	10,81	27,18
Services emp. r.	27	77,03	78,24	4,50	0,06	18,68	69,83	88,51

Table A5: Descriptive statistics in group 5

Group 5	N	Mean	Median	Std. Deviation	Coefficient of variation	Range	Minimum	Maximum
Incidence	15	278.52	290.80	32.06	0.12	117.8	220.4	338.2
Mortality	15	127.27	124.50	13.33	0.10	45.5	106.2	151.7
5-year prevalence	15	1561.53	1550.80	335.79	0.22	1133.5	904.6	2038.1
GDP per capita	15	14786.33	15653.56	6566.13	0.44	19297.7	5913.0	25210.7
HDI	15	0.85	0.85	0.04	0.05	0.1	0.8	0.9
Fertility rate	15	1.55	1.55	0.12	0.08	0.5	1.3	1.8
Trade to GDP rate	15	115.44	108.43	32.27	0.28	101.7	69.3	171.0
Urban population ratio	15	63.21	60.04	9.41	0.15	30.4	48.6	79.0
Rate of agricultural employment	15	8.87	6.62	5.85	0.66	18.6	2.7	21.2
Rate of industrial employment	15	30.55	30.38	3.65	0.12	13.5	23.7	37.3
Rate of services employment	15	60.58	61.12	6.13	0.10	20.3	48.7	69.0

A.2 Correlation matrices per group

Table A6: Correlation matrix in group 1

Group 1	Inci-dence	Morta-lity	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.944	0.785	0.096	0.186	-0.310	0.009	-0.201	0.319	-0.314	-0.271
Mortality		1	0.612	-0.078	-0.031	-0.216	-0.047	-0.242	0.418	-0.414	-0.354
5-year prevalence			1	0.422	0.583	-0.669	0.198	-0.047	-0.005	0.022	-0.003
GDP per capita				1	0.633	-0.225	0.227	0.549	-0.333	0.244	0.318
HDI					1	-0.552	0.138	0.375	-0.442	0.408	0.387
Fertility r.						1	-0.260	0.011	0.276	-0.231	-0.252
Trade to GDP r.							1	0.177	-0.205	0.190	0.178
Urban pop. r.								1	-0.589	0.284	0.628
Agric. emp. r.									1	-0.736	-0.956
Industrial emp. r.										1	0.505
Services emp. r.											1

Table A7: Correlation matrix in group 2

Group 2	Incidence	Mortality	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.713	0.765	-0.223	0.180	-0.043	-0.050	0.102	-0.117	-0.037	0.161
Mortality		1	0.325	-0.397	-0.241	0.099	0.131	-0.151	0.227	-0.038	-0.245
5-year prevalence			1	-0.145	0.415	-0.367	-0.151	0.077	-0.070	-0.094	0.140
GDP per capita				1	0.653	-0.309	0.048	0.534	-0.563	0.645	0.274
HDI					1	-0.368	-0.187	0.566	-0.632	0.460	0.468
Fertility r.						1	-0.039	-0.004	-0.087	-0.042	0.128
Trade to GDP r.							1	0.049	-0.086	0.011	0.094
Urban pop. r.								1	-0.622	0.308	0.548
Agric. emp. r.									1	-0.534	-0.857
Industrial emp. r.										1	0.023
Services emp. r.											1

Table A8: Correlation matrix in group 3

Group 3	Incidence	Mortality	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.418	0.762	0.403	0.509	-0.264	0.013	0.207	-0.235	0.131	0.162
Mortality		1	-0.023	-0.297	-0.238	0.250	0.327	-0.258	0.399	0.223	-0.447
5-year prevalence			1	0.530	0.775	-0.640	-0.183	0.253	-0.429	0.028	0.377
GDP per capita				1	0.771	-0.240	-0.152	0.355	-0.577	-0.109	0.564
HDI					1	-0.468	-0.166	0.437	-0.475	0.042	0.413
Fertility r.						1	0.152	-0.233	0.369	-0.068	-0.307
Trade to GDP r.							1	-0.204	0.288	0.093	-0.296
Urban pop. r.								1	-0.451	-0.043	0.424
Agric. emp. r.									1	0.047	-0.923
Industrial emp. r.										1	-0.428
Services emp. r.											1

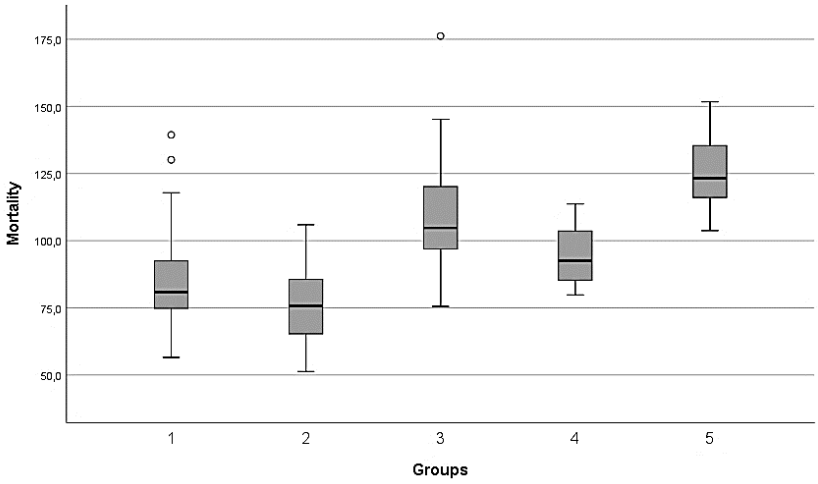
Table A9: Correlation matrix in group 4

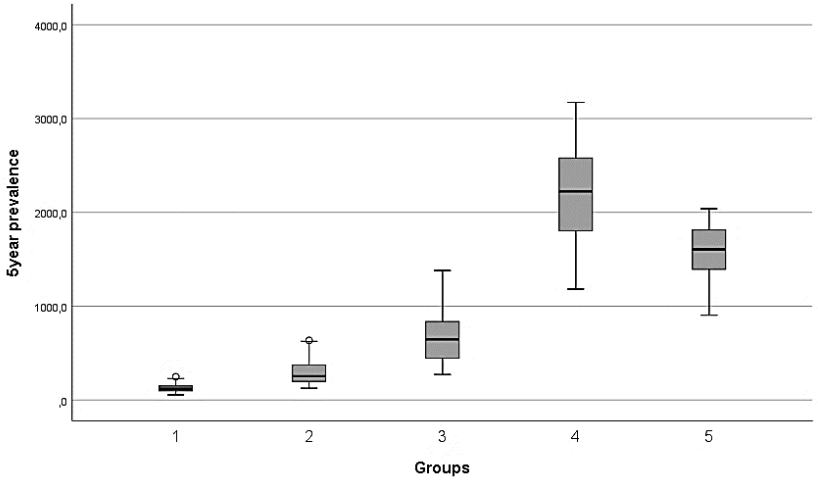
Group 4	Incidence	Mortality	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.129	0.889	0.152	0.347	0.686	-0.220	-0.074	-0.038	-0.017	0.035
Mortality		1	-0.025	-0.107	0.025	0.162	0.462	-0.052	0.117	-0.153	0.064
5-year prevalence			1	0.050	0.374	0.611	-0.255	-0.117	-0.083	0.204	-0.125
GDP per capita				1	0.539	0.189	0.093	-0.023	-0.378	-0.421	0.561
HDI					1	0.459	0.210	0.040	-0.350	-0.045	0.230
Fertility r.						1	-0.300	-0.111	-0.064	0.009	0.028
Trade to GDP r.							1	0.334	-0.324	-0.181	0.330
Urban pop. r.								1	-0.341	-0.476	0.587
Agric. emp. r.									1	-0.007	-0.545
Industrial emp. r.										1	-0.835
Services emp. r.											1

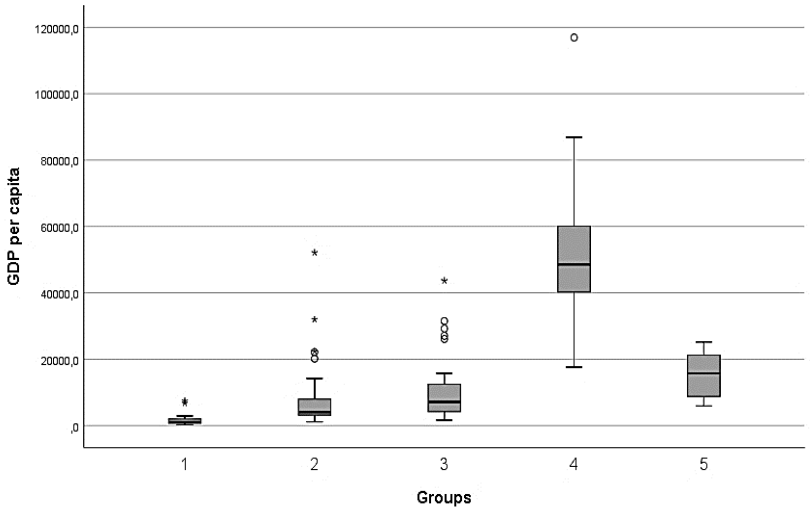
Table A10: Correlation matrix in group 5

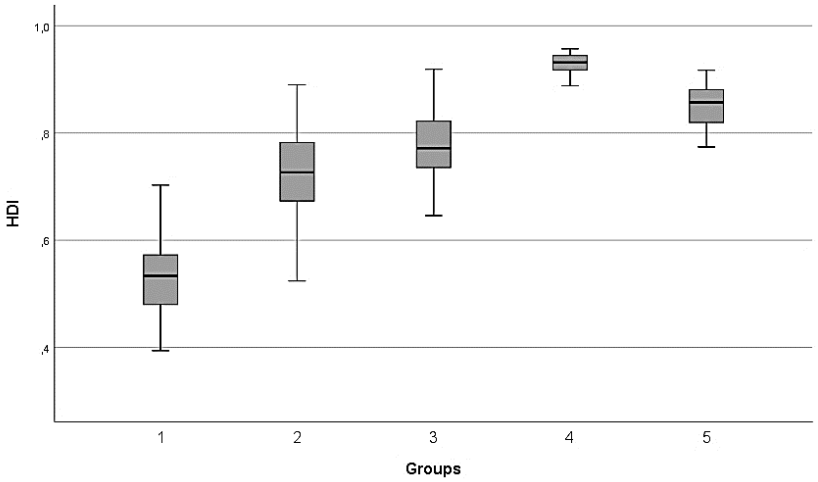
Group 5	Incidence	Mortality	5-year prevalence	GDP per capita	HDI	Fertility r.	Trade to GDP r.	Urban pop. r.	Agric. emp. r.	Industrial emp. r.	Services emp. r.
Incidence	1	0.416	0.854	0.576	0.646	0.412	0.408	0.136	-0.538	0.042	0.457
Mortality		1	-0.019	-0.319	-0.222	-0.139	-0.021	-0.362	0.252	0.052	-0.257
5-year prevalence			1	0.850	0.890	0.489	0.309	0.269	-0.734	-0.013	0.666
GDP per capita				1	0.934	0.507	0.326	0.134	-0.740	0.073	0.620
HDI					1	0.515	0.340	0.243	-0.741	0.124	0.590
Fertility r.						1	0.281	0.275	-0.233	0.112	0.142
Trade to GDP r.							1	0.136	-0.522	0.654	0.075
Urban pop. r.								1	-0.462	-0.134	0.496
Agric. emp. r.									1	-0.155	-0.804
Industrial emp. r.										1	-0.462
Services emp. r.											1

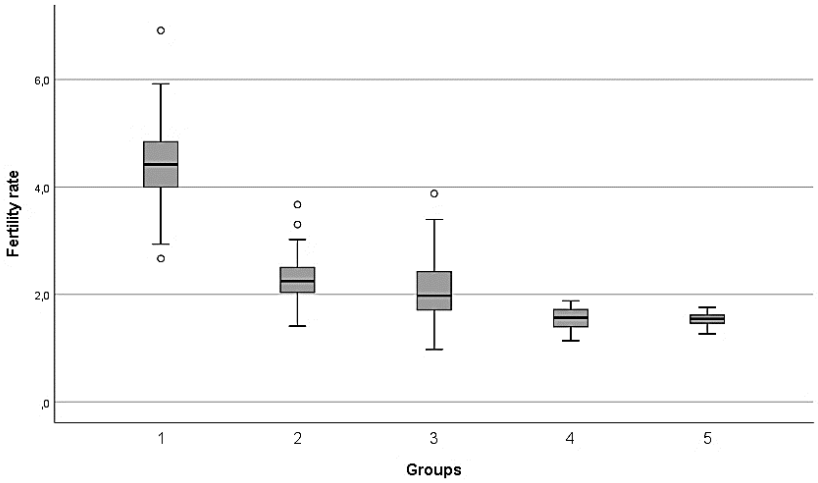
A.3 Box-plots

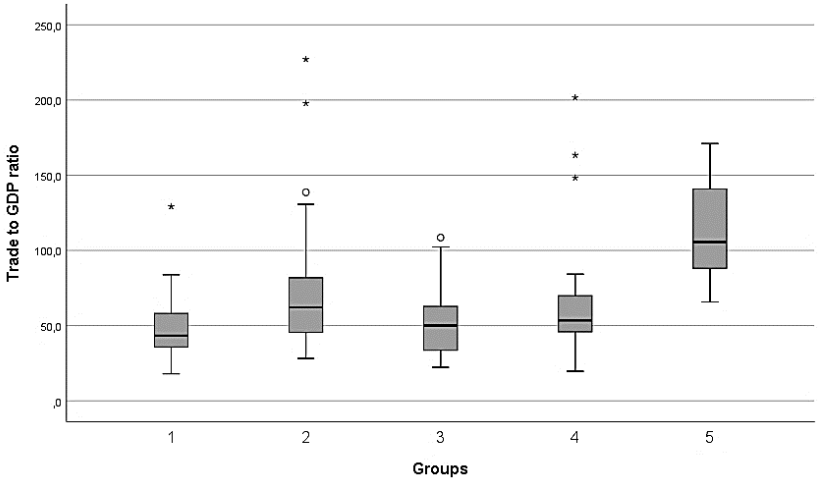

175,0
150,0
125,0
100,0
75,0
50,0
Mortality
1
2
3
4
5
Groups

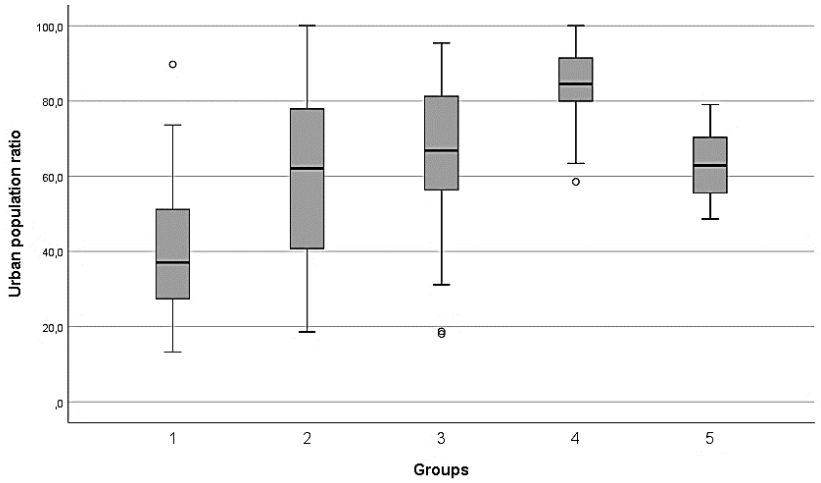

4000,0
3000,0
2000,0
1000,0
,0
5year prevalence
1
2
3
4
5
Groups

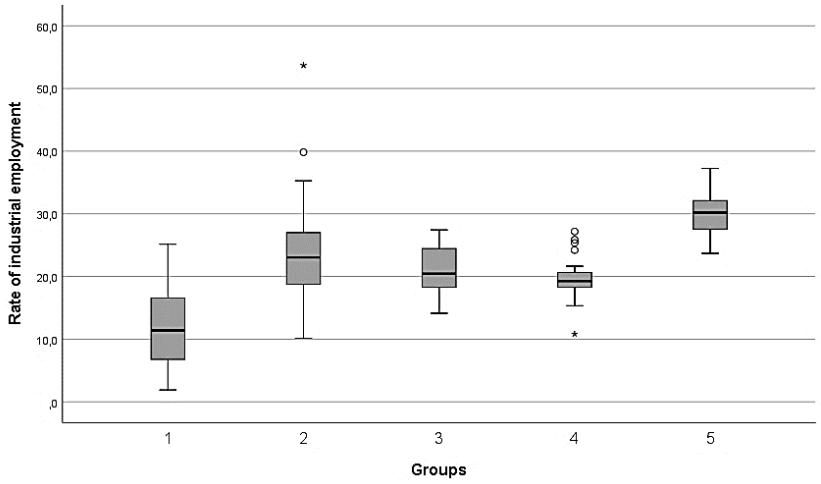

GDP per capita
120000,0
100000,0
80000,0
60000,0
40000,0
20000,0
,0
1
2
3
4
5
Groups

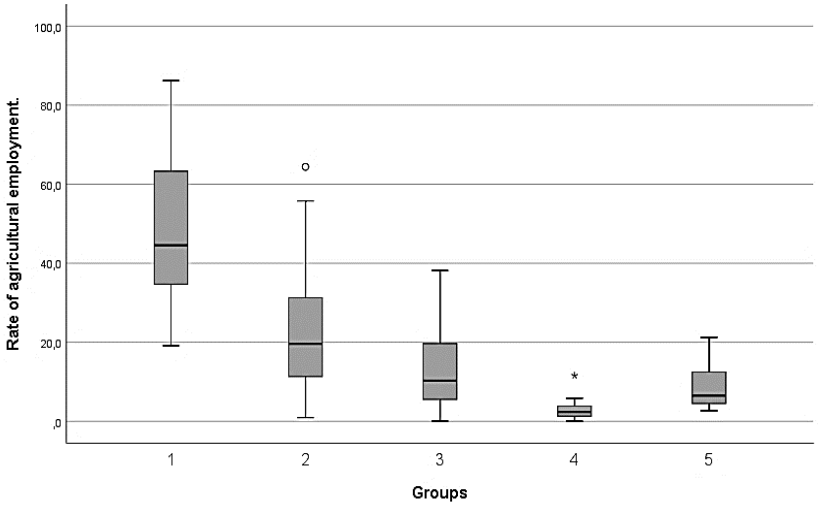

HDI
1,0
,8
,6
,4
1
2
3
4
5
Groups

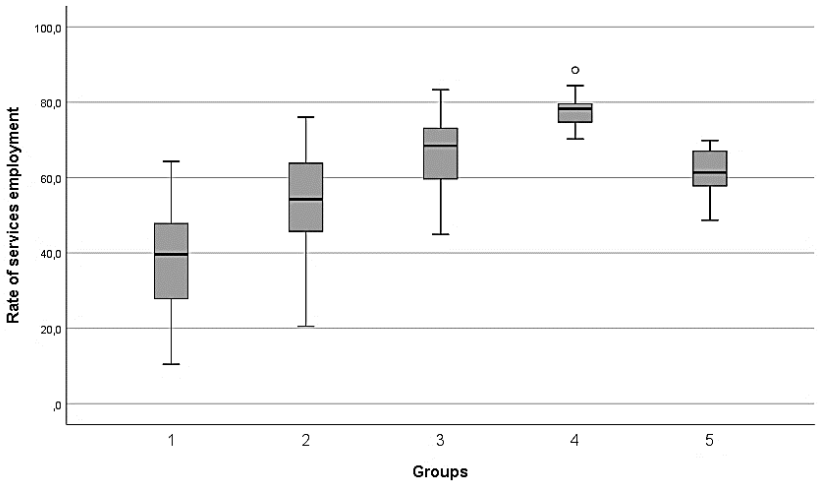

Fertility rate
6,0
4,0
2,0
,0
1
2
3
4
5
Groups


Trade to GDP ratio
250,0
200,0
150,0
100,0
50,0
,0
1
2
3
4
5
Groups


Urban population ratio
100,0
80,0
60,0
40,0
20,0
,0
1
2
3
4
5
Groups


Rate of industrial employment
60,0
50,0
40,0
30,0
20,0
10,0
,0
1
2
3
4
5
Groups


Rate of agricultural employment
100,0
80,0
60,0
40,0
20,0
,0
1
2
3
4
5
Groups


Rate of services employment
100,0
80,0
60,0
40,0
20,0
,0
1
2
3
4
5
Groups