Does Economic Status Matter for the Regional Variation of Malnutrition-Related Diabetes in Romania? Temporal Clustering and Spatial Analyses

Druica, Elena and Goschin, Zizi

University of Bucharest, Romania, The Bucharest University of Economic Studies, Romania

2016

Does Economic Status Matter for the Regional Variation of Malnutrition-Related Diabetes in Romania?
Temporal Clustering and Spatial Analyses

Elena Druică
Faculty of Business and Administration,
University of Bucharest, Romania

Email: elena.druica@faa.unibuc.ro

Zizi Goschin
The Bucharest University of Economic Studies, Romania
Institute of National Economy, Romanian Academy

E-mail: zizi.goschin@csie.ase.ro

Abstract

Among the different types of diabetes, in Romania the malnutrition-related diabetes displays the highest territorial inequality. In this paper, we combined two types of statistical tools, temporal clustering and spatial analysis, to find some relevant patterns in its territorial distribution. Firstly we conducted a time series clustering for the 41 counties and Bucharest Municipality, over 2007-2014, based on CORT and ACF dissimilarity distances and choose four clusters in each case. Within each cluster the evolution of malnutrition diabetes is similar. The clusters were then included as dummy variables in a spatial model testing the determinants of malnutrition-related diabetes incidence at county level. Malnutrition-related diabetes is a disease that might be linked to the economic status, therefore GDP per capita and average wage have been tested and found significant as factors of influence in various model specifications. The dummies representing the temporal clusters are also significant determinants of the regional incidence of malnutrition-related diabetes in Romania. We found that when introducing the cluster dummies in the spatial model, it becomes less appropriate than classic OLS regression, which suggests that temporal clusters were able to capture the spatial dependence in our data. The contribution of our work is three folded. First, we applied time series clustering in R and in doing so we added a real – data application to this scarce stream of literature. Secondly, we combined two techniques relatively new in Romanian data: spatial analysis and time series clustering. Last, but not least, we discussed the malnutrition – related diabetes, mellitus, as a possible proxy of poverty, and tried to advocate our claim by relating this disease's territorial distribution with some economic variables.

Key words: malnutrition-related diabetes, time series clustering, spatial clustering, spatial-lag model, county.

1. Introduction

The departure point for our investigation is the fact that, according to The National Center for Statistics and Informatics for Public Health (CNSISP), "diabetes prevalence in

Romania varies by region" (Chiriac & Scorțan, 2015), as a result of different factors among which economic status can play an important role.

The focus of this paper is a rare type of diabetes associated with long-term malnutrition, namely malnutrition-related diabetes – mellitus. This is a type of diabetes that can be linked to economic variables (like living standards) more than other types. A large body of literature is devoted to diabetes driven by long-term malnutrition in the developing countries, but there are also some contributions pointing to a new type of malnutrition, more specifically modern malnutrition, that is more common in people from lower socioeconomic groups even in developed countries (see the online reference). In this case it is not the chronic under nutrition due to lack of food, specific to the underdeveloped world, but rather a nutrient deficiency generated by poor diet habits, such as large consumption of low-quality and protein deficient fast food, excess of fats and sugar in the diet, etc. Although the incidence of this type of diabetes tends to decrease, the high inequalities in its distribution across Romania, and its presumed link with poverty effects, invite for analysis.

In this context, our paper explores the characteristics of regional variability in malnutrition-related diabetes in Romania, in two conceptually different but complementary ways, specifically temporal and spatial clustering. Both methods are appropriate for our research endeavor, and have high explanatory power, especially when used together. Temporal clustering identifies the common regional trends in the dynamics of new cases of malnutrition-related diabetes, while spatial clustering reveals the concentration of similar levels in space. By applying both methods, we can approach this topic in a more comprehensive way, enabling us to achieve a deeper understanding of the specific temporal and spatial patterns of the malnutrition-related diabetes in Romania. To the authors' best knowledge, this is the first study to combine temporal and spatial clustering. Moreover, we make an attempt to include the common regional trends identified through temporal clustering into both classic and spatial regression models, using cluster dummies.

The research is conducted in three stages: in the first stage, based on GINI index calculation, we pinpoint that among the nine categories of diabetes for which the Romanian Institute of Public Health provides data, the malnutrition diabetes has the highest inequality in distribution across Romania. In the second stage we apply a time series cluster analysis to find groups of counties within malnutrition diabetes that have similar trend patterns. Third, a parallel spatial clustering is performed to highlight the spatial dependence in the territorial distribution of this disease. Finally, the temporal clusters are used to develop a territorial analysis in the framework of regression models.

The remainder of this paper proceeds as follows. Next section reviews the sparse international literature on malnutrition-related mellitus diabetes. Section 3 outlines the methodological framework of our research. Section 4 displays the outcomes from our four lines of research and discusses both their statistical significance and economic implications. Finally, section 5 concludes by summarizing the main results and tracing directions for further research.

2. Literature review

There is a small body of literature addressing the malnutrition-related mellitus diabetes: it seems to be a rare type of diabetes associated with long-term malnutrition, one of the reasons why the literature concerning the developed part of the world is almost missing, while it is of interest in the developing countries: see for example (Lester, 1993). On the other hand, there are contributions discussing the mellitus diabetes in relation to obesity, a common problem in developed countries (George, Jacob & Fogelfeld, 2015).

The World Health Organization reports two types of malnutrition-related diabetes:

one that is characterized by a socioeconomic setting of poverty and malnutrition, onset in youth mainly below the age of 30 years, and another category of protein-deficient pancreatic diabetes which has similar characteristics as the previous one, but differs in absence of several clinical and radiological evidence of pancreatic dysfunction and relative resistance to insulin (WHO report, 1985). Although the CNSISP, the source of our data, does not provide a definition of this type of diabetes, we are told that the number of cases in this category halved between 2007 and 2013, a result that holds both by gender and by residence (Chiriac & Scorțan, 2015).

Although of less interest in health public policy, as other conditions are claiming for more interest and more urgent interventions, malnutrition diabetes is considered as linked to poverty (Taksande et.al, 1985). Most likely too weak to be considered a proxy for the phenomenon as a whole, we rely on the assumption that we can still use the territorial distribution of this disease in relation to living standards. The World Health Organization acknowledges the need for assessing the distribution of health risks by socioeconomic position at national and local levels (Blakely T, Hales S, Woodward A. 2004) and UNICEF provides studies regarding effective ways to target poor people in developing countries, among which proxy means testing has become increasingly popular (Kidd & Wylde, 2011). Several common proxies have been discussed in previous research, like malaria (Worrall, Basu & Hanson, 2003), other tropical diseases (Samuels & Pose, 2013), or the worm index (Hotez & Herricks, 2015), to name only a few. However, these contributions address the developing countries and none of the diseases considered are relevant for countries that reached at least middle - income level.

Proxies for poverty are less discussed for moderately developed, and even less for developed countries, although disparities in health are commonly recognized in relation to poverty. Our study focuses on malnutrition diabetes and investigates whether the incidence of this condition is by any means related to several relevant economic indicators.

3. Methodological framework

The data set we used is provided by CNSISP, cover a time span between 2007 and 2014, are annual data, and include nine categories of diabetes: incidence and prevalence of diabetes recorded by nutritionists; incidence and prevalence of diabetes recorded by family doctors, as well as the incidence of the following categories: malnutrition diabetes; insulin diabetes; non – insulin diabetes; unspecified, and other types of diabetes. Data were collected by each Romanian county and Bucharest Municipality, 42 NUTS3 statistical units in total, and analyzed for 8 years. The variables are reported as indices per 100000 inhabitants.

The GINI index. The method we choose to depart from is suggested by Schneider (Schneider et. al, 2001), and was previously applied in analyzing the inequality in infant mortality rate. The first step is to calculate the GINI index for our variable of interest, an then to apply a cluster analysis and find similar groups of statistical units. As the authors pinpoint, the Lorenz curve is usually related to the assessment of income inequality distribution, but in general terms it is nothing but a cumulative frequency distribution that aims to compare a certain distribution with the uniform distribution of complete equality. The GINI coefficient calculated based on Lorentz curve is the ratio of two areas. One is the area between the line of perfect equality and the Lorenz curve, while the other one is the area between the line of perfect equality and the one of perfect inequality. In short, for a discrete distribution, GINI index can be calculated according to the following formula:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i} \qquad (1)$$

Due to the special nature of the type of diabetes we discuss, malnutrition – related, our first goal is to see whether its territorial distribution differs by comparison with other types.

Time series clustering. In the next stage, a time series clustering is conducted, using R software and two dedicated packages: "clustertend" and "TSclust". The first package was used to evaluate the clustering tendency based on Hopkins statistic; the second was used to get the groups.

The cluster analysis is meant to extract groups of counties similar in their incidence of malnutrition diabetes. Since the available data were 42 time series for a time span of 8 years, the tools specific to cluster analysis for cross – sectional data are not sufficient anymore, as they don't capture either dynamics, or autocorrelation in data (Liao, 2005). An extensive body of literature has been developed recently for time series clustering, along with software able to handle this type of analysis, with the main focus on clarifying the notions of similarity and dissimilarity between time series, rather than on developing new clustering methods (Paparrizos & Gravano , 2015). The main R package we use, TSclust, provides four types of distance categories, namely: model – free distance, model – based distance, complexity – based distance and predicted – based distance, previously documented in the literature (Liao, 2005). However, there is no agreement among specialists what distance to choose for measuring the proximity, this choice largely depending on data at hand (Montero&Villar, 2014).

After analyzing the clustering tendency of our data based on "clustertend" package, we found that the value of Hopkins statistic is H = 0.22, which is far below the threshold 0.5. Therefore, we concluded that the dataset is highly clusterable (Banerjee & Dave, 2004). Next, we decided to work first with the CORT distance, a model – free proximity measure that combines two characteristics. On the one hand, it measures the closeness of the time series values at different points in time based on conventional Euclidean distance. On the other hand, it takes into consideration a temporal correlation of first order. The CORT distance is therefore a dissimilarity measure that covers both conventional measures for the proximity on observations as well as temporal correlation (Montero&Villar, 2014), and has the following formula:

$$d_{CORT} = \Phi_k(CORT) * d_{XY} \quad (2)$$

where d_{XY} is the Euclidean distance between the time series X and Y, namely

$$d_{XY} = \sqrt{\sum_{t=1}^{T}(X_t - Y_t)^2} \quad (3)$$

CORT is the temporal correlation coefficient of order one:

$$CORT(X,Y) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2 \sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2} \quad (4)$$

and $\Phi_k()$ is a modulating function having the formula:

$$\Phi_k(x) = 2/(1+\exp(kx)),\ k > 0 \quad (5)$$

In TSclust, there is a default setting for k, k = 2 and we kept this value in our analysis. The choice of this particular distance is justified by its double capacity to account for temporal and spatial characteristic in data. In the meanwhile, more complicated distances,

like for example model – based distances, may not be appropriated since the time series are short.

The second distance we considered is the ACF distance, which compares sequences of serial correlation extracted from the original time series (Galeano & Pena, 2000). It can be calculated based on the following formula:

$$d_{ACF}(X,Y) = \sqrt{(\hat{\rho}_X - \hat{\rho}_Y)^T \Omega (\hat{\rho}_X - \hat{\rho}_Y)} \qquad (6)$$

where Ω is a matrix of weights, and $\hat{\rho}_X$ and $\hat{\rho}_Y$ are the estimated autocorrelation vectors of the time series X and Y. The default lag considered in this calculation by TSclust is 50, which means that for our data covering only 8 years, there is no risk to miss a significant lag. As can be observed, if the weight matrix is the unit matrix, which is the case in our analysis, then the ACF distance is the Euclidean distance between the autocorrelation functions of X and Y. Although there is no clear interpretation of this distance in the literature, we considered it as a measure of similarity in the "one moment to another memory".

The clustering algorithm used in our investigation is hierarchical clustering, accepted not only for cross – sectional data but also for time series (Kaufman & Rousseeuw, 2005). Unfortunately, at the time being and to our knowledge, there is no package in R able to run cluster validation for time series clustering. Several packages are used for this purpose, like for example "clusterSim", or "fpc" among others, but they consider in essence only cross – sectional data. The limits of their applicability in time series clustering come from the special dissimilarity distances involved in the later case, which are not included in the mentioned, or in other packages.

The "TSclust" package provides, though, a hint toward validation: a function named cluster.validation(), which is recommended for evaluate the similarity between two clusters (Montero&Villar, 2014). The illustration of how this function works is however presented as a comparison between the clustering result and the "true cluster", which means that at this stage it can only work for experimental data, when the true cluster is known. Another recommendation is to validate clusters through their practical meaning and interpretation.

Spatial clustering. Space is largely acknowledged as an important factor of influence for many diseases and in order to achieve a better understanding of the regional distribution of diabetes and to uncover its possible causal factors, it is useful to establish its precise distribution in space. This implies highlighting the main spatial concentration patterns of malnutrition-related diabetes by finding its spatial clusters and outliers.

The first option is the traditional Moran's *I* indicator (Anselin and Rey, 1991) that allows us to measure the global spatial dependence in malnutrition-related diabetes (MD), as follows:

$$MI = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (MD_i - \overline{MD})(MD_j - \overline{MD})}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (MD_i - \overline{MD})^2} \qquad (7)$$

where MD_i and MD_j represent the values of the malnutrition-related diabetes incidence in the counties *i* and *j* respectively, $\overline{MD}$ is the national average, and w_{ij} represent spatial weights capturing the "spatial influence" between county *j* and county *i*. Moran's *I* requires that the local neighborhood around each geographic unit is defined based on a weights matrix. In this paper we test first-, second- and third-order queen contiguity matrices, i.e. w_{ij} = 1 if regions *j and i* are neighbors and w_{ij} = 0 otherwise.

Spatial autocorrelation can be approached from two perspectives: global or local. Global indicators such as Moran capture the whole area through a single synthetic value. It is useful to measure the spatial association for each individual location *i* as well, as it might point to local specifics and spatial clusters and outliers. To this end we use the LISA (Local Indicators of Spatial Autocorrelation) indicator, defined as follows (Anselin, 2005; LeSage and Pace, 2009):

$$LISA_i = Z_i \sum_{j=1}^{m} w_{ij} Z_j \quad (8)$$

where z_i and z_j are the standardized scores of malnutrition-related diabetes incidence in the counties *i* and *j* respectivelly, *j* representing only the neighbors of county *i* (as defined by the weights w_{ij}). The Cluster Map associated to the Local Indicators of Spatial Autocorrelation in Geoda points to significant cases of local spatial dependence by type of spatial correlation (positive-similar or negative-dissimilar). LISA mapping allows identification of local deviations that are significant for the territorial distribution, providing better understanding of spatial processes (Fotheringham, 1997; Unwin and Unwin, 1998).

Spatial models. Aiming to test the determinants of MD in Romania, we will be further using specific methods of spatial analysis.

Firstly, a classic regression model is employed for estimating the influence of various likely determinants of regional malnutrition-related diabetes in Romania, as follows:

$$MD_i = a + \sum_k b_k X_{ki} + \varepsilon_i \quad (9)$$

where X_k are the regressors and $\boldsymbol{\varepsilon}$ is the error term. Given that economic status can play an important role in the emergence and evolution of malnutrition-related diabetes, we are going to test GDP per capita (proxy for county development) and average wage as potential factors of influence. A one-year lag (MD_{t-1}) of the dependent variable MD is also included as explanatory variable to account for the potential inertia of this phenomenon.

If spatial dependence is confirmed, it should be corrected using the appropriate spatial model (Anselin, 2005; LeSage and Pace, 2009). To this aim two main types of spatial models are going to be tested: spatial lag and spatial error model. The first one is a spatially autoregressive model, including the spatial lag of the dependent variable ($\rho \sum_j w_{ij} MD_j$) in the previous classic model specification:

$$MD_i = a + \sum_k b_k X_{ki} + \rho \sum_j w_{ij} MD_j + \varepsilon_i \quad (10)$$

while the spatial error model accounts for spatial dependence through its spatially autoregressive error term, as follows:

$$MD_i = a + \sum_k b_k X_{ki} + (\lambda \sum_j w_{ij} \varepsilon_j + \upsilon_i) \quad (11)$$

We will finally choose the most appropriate model for our data according to the value of Lagrange multiplier test and other statistics.

This paper not only employs temporal and spatial clustering to analyze the same dataset, but also explores the possibility to gather the information on common regional trends, identified through temporal clustering, in the spatial regression models. To this end, the temporal clusters determined at an earlier stage of the analysis will be introduced as dummies in the previous model specifications in order to test their relevance. For the classic OLS model, the new specification is as follows:

$$MD_i = a + \sum_k b_k X_{ki} + \sum_m CL_m + \varepsilon_i \quad (12)$$

where CL_m represent the dummy variables identifying each county's inclusion in one of the four temporal clusters. Similarly, the dummy variables are included in spatial equations (10) and (11). Thus, the extended spatial models allow us to capture the spatial dependence in the territorial distribution of malnutrition-related diabetes, while harvesting the information previously provided by time clusters by including them explicitly in the models.

4. Results and discussions

In the first step of our analysis we calculated the GINI index for all 9 categories of diabetes recorded by the Romanian National Institute of Public Health, and compared the results. Table 1 shows the evolution of GINI index between 2007 and 2014, calculated for 42 statistical units, covering all Romanian counties and the capital, Bucharest Municipality. The malnutrition-related diabetes shows the highest fluctuations, as well as the highest values among the nine categories, suggesting that it follows different patterns and most likely it is influenced by factors that are not necessary relevant for the rest.

Table 1: GINI Indexes for Romania, 2007 – 2014, for different types of diabetes

Category	2007	2008	2009	2010	2011	2012	2013	2014
Nutritionists – prevalence	0.23	0.22	0.22	0.19	0.21	0.21	0.17	0.23
Nutritionists - incidence	0.23	0.17	0.19	0.18	0.21	0.26	0.31	0.23
Family doctors - prevalence	0.23	0.18	0.21	0.17	0.21	0.21	0.19	0.22
Family doctors - incidence	0.28	0.16	0.18	0.26	0.27	0.28	0.17	0.28
Malnutrition - incidence	**0.60**	**0.49**	**0.43**	**0.30**	**0.70**	**0.48**	**0.33**	**0.63**
Insulin - incidence	0.38	0.31	0.23	0.27	0.22	0.27	0.21	0.38
Noninsulin - incidence	0.28	0.18	0.19	0.26	0.29	0.30	0.18	0.42
Other types, incidence	0.36	0.48	0.22	0.34	0.63	0.39	0.51	0.70
Unspecified, incidence	0.46	0.48	0.36	0.20	0.46	0.57	0.23	0.67

Source: own processing

We further attempt to find potential patterns – both in time and space – for the incidence of this disease, and then we look for some economic factors that may explain the high inequality in its territorial distribution.

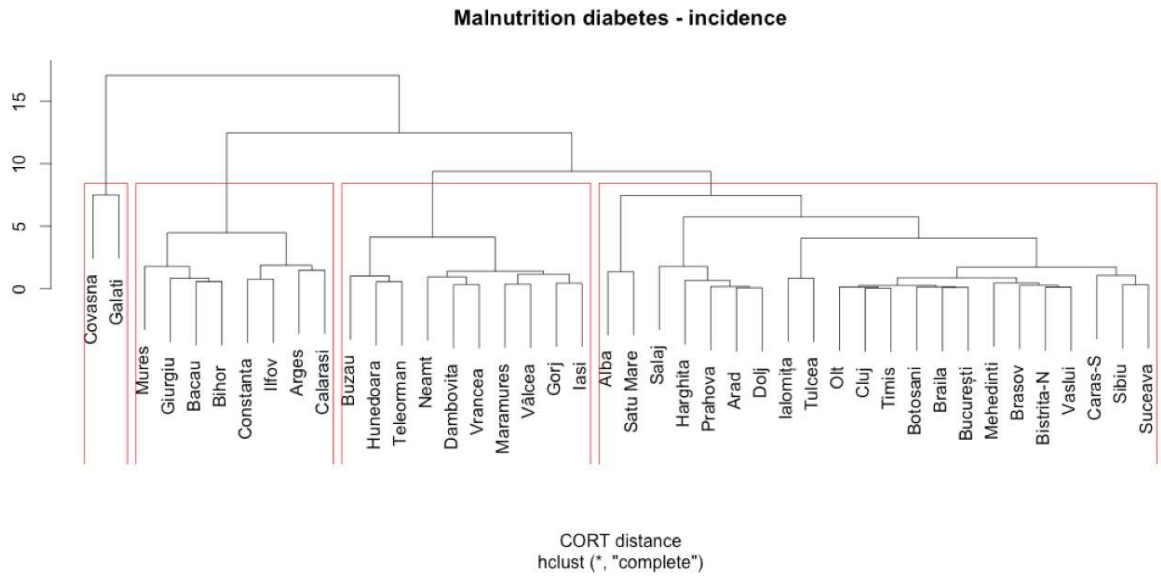
In Table 2 we report the results of the time series clustering based on CORT distance and ACF distance respectively. The territorial distribution of the clusters is provided in Figure 2 and 3. The clusters created are different, but some groups still share some common counties. Some differences between clusters were expected, since the two distances measure different things. In the first case, the CORT distance points toward counties that share similarity in shape, while the ACF distance seems to be a measure of time series memory from one year to another.

Table 2: Clusters based on CORT and ACF distances, respectively

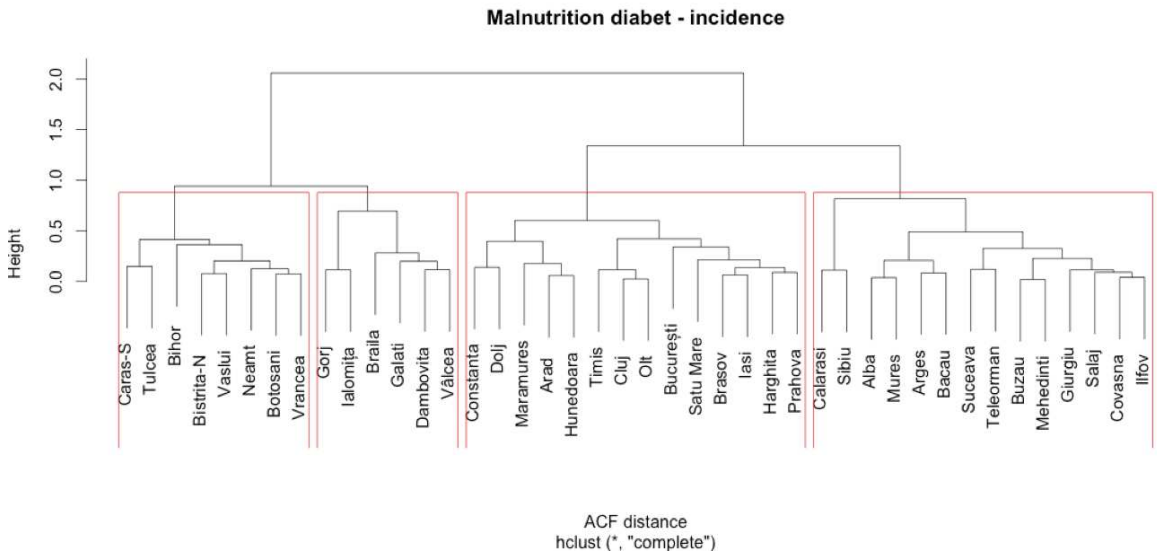
Distance	Group 1	Group 2	Group 3	Group 4
CORT	Alba, Arad, Bistrita-Nasaud, Botosani, Brasov, Braila, Caras-Severin, Cluj, Dolj, Harghita, Ialomița, Mehedinti, Olt, Prahova, Satu Mare, Salaj, Sibiu, Suceava, Timis, Tulcea, Vaslui, București	Arges, Bacau, Bihor, Calarasi, Constanta, Giurgiu, Ilfov, Mures	Buzau, Dambovita, Gorj, Hunedoara, Iasi, Maramures, Neamt, Teleorman, Vâlcea, Vrancea	Covasna, Galati
ACF	Alba, Arges, Bacau, Buzau, Calarasi, Covasna, Giurgiu, Ilfov, Mehedinti, Mures, Salaj, Sibiu, Suceava, Teleorman	Arad, Brasov, Cluj, Constanta, Dolj, Harghita, Hunedoara, Iasi, Maramures, Olt, Prahova, Satu Mare, Timis, București	Bihor, Bistrita-Nasaud, Botosani, Caras-Severin, Neamt, Tulcea, Vaslui, Vrancea	Braila, Dambovita, Galati, Gorj, Ialomița, Vâlcea

Source: own processing

The dendograms are presented in Figure 1 (a) and (b), while the average values within each group are recorded in Table 3 (for CORT distance) and Table 4 (for ACF distance).



(a)



(b)

Figure 1: The dendograms based on CORT distance (a) and ACF distance (b); hierarchical clustering

Source: own processing

The first group presented in Table 3 includes the majority of counties, 52% of them, and shows an abrupt drop from a rather high incidence, of 13.73, to small values, which are maintained over the entire period of time.

Table 3: The average incidence values within each group, CORT distance

	2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**prop**
Group 1	13.73	2.82	2.56	1.24	2.25	1.38	1.84	1.31	**0,52**
Group 3	8.74	7.80	4.95	6.89	5.85	7.13	1.46	2.78	**0,24**
Group 2	3.15	10.82	10.29	11.39	3.94	1.20	3.21	7.14	**0.19**
Group 4	14.41	62.85	23.20	3.83	18.07	5.55	15.93	11.45	**0.05**

The highest overall incidence corresponds to the fourth group that includes only two counties. With a moderate average incidence level, Group 3 includes 24% of the counties where the incidence remained high until 2012, followed by a significant drop that is maintained until 2014. Group 4 seems to have the most irregular behavior among the four, alternating from periods with high incidence to periods with low incidence. The evolution of these four groups is presented graphically in Figure 2.

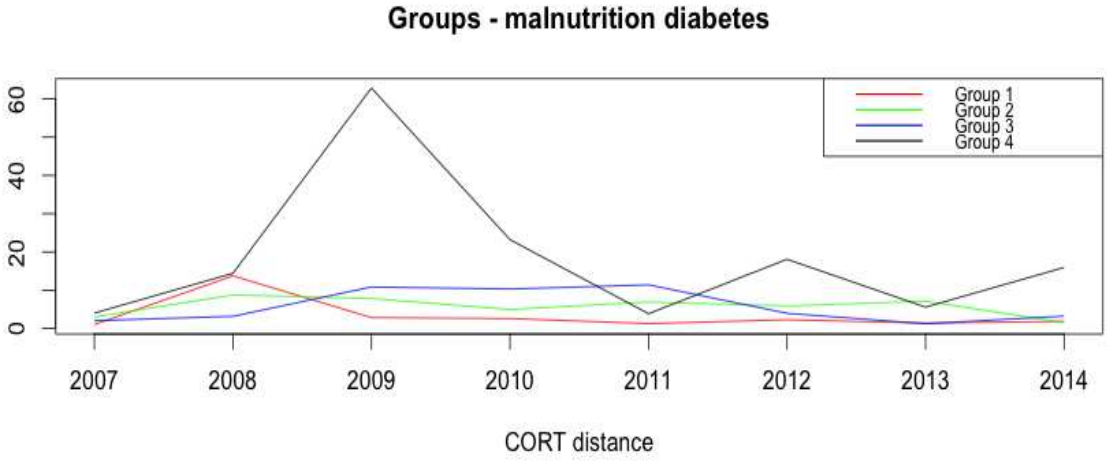


Figure 2: Incidence of malnutrition diabetes – evolution in time; CORT distance
Source: own processing

The next table includes the information regarding clusters produced based on the ACF distance. Unlike the CORT distance, in this case the four groups are more balanced in size: the first and second groups include 33% of the counties each, while the third and the fourth comprise of 19% and 14% respectively.

Table 4: The average incidence values within each group, ACF distance

	2007	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**prop**
Group 1	11.52	16.25	9.02	6.47	5.30	4.00	2.77	5.01	**0,33**
Group 2	12.94	2.84	2.55	2.72	2.21	1.95	1.10	1.98	**0,33**
Group 3	6.35	4.55	4.24	4.14	2.82	2.08	3.29	1.57	**0.19**
Group 4	8.38	8.07	6.44	5.54	8.03	3.74	5.35	4.36	**0.14**

As the graphical presentation in Figure 2 shows, the average incidence within the four groups obtained based on the ACF distance is more variable than in the previous case, presented in Figure 3.

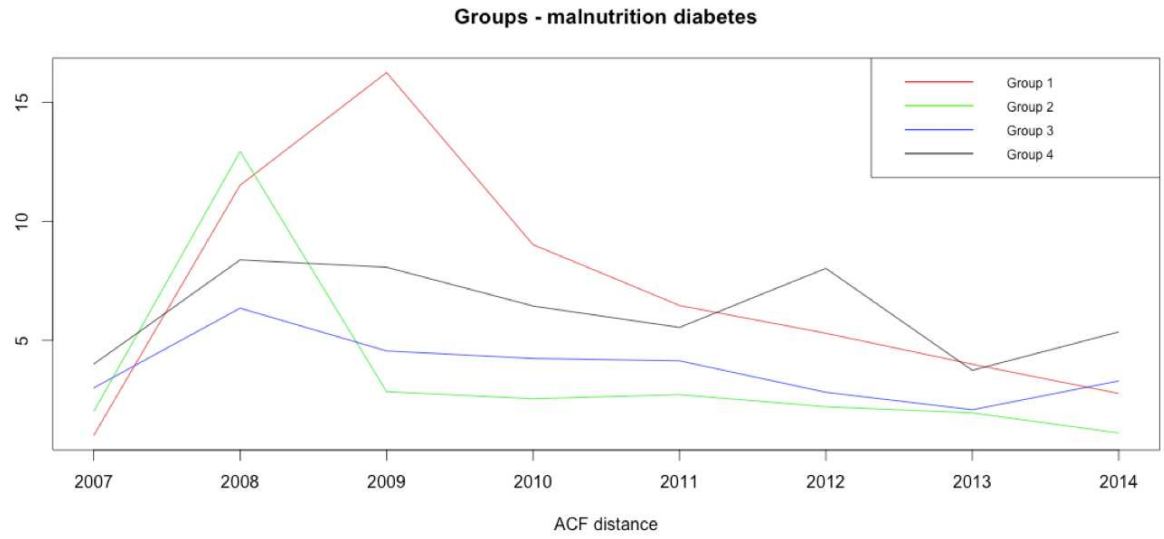
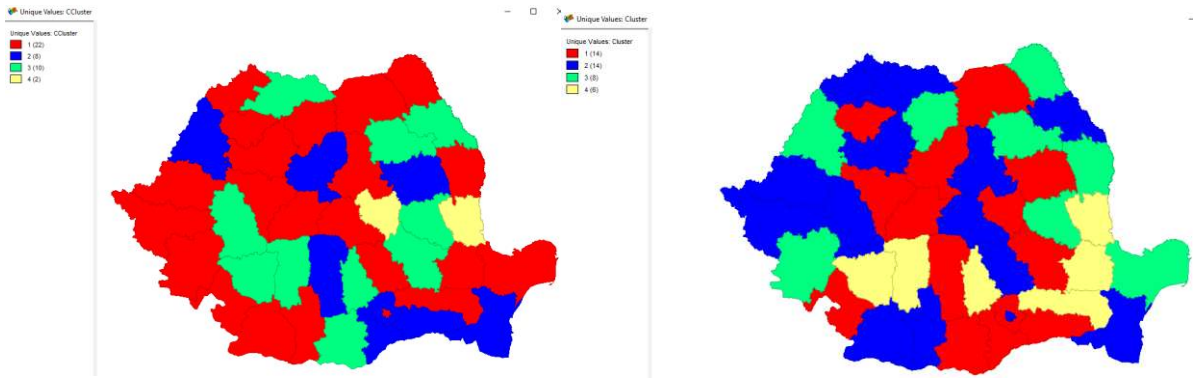


Figure 3: Incidence of malnutrition diabetes – evolution in time; ACF distance
Source: own processing

The CORT distance yielded to three groups very similar and a fourth one very different, which creates a first expectation that in a regression model only Group 4 will be statistically significant in respect to the reference group. For the ACF distance however, the groups display different profiles and we expect at least Group 1 and 4 to differ form the reference.



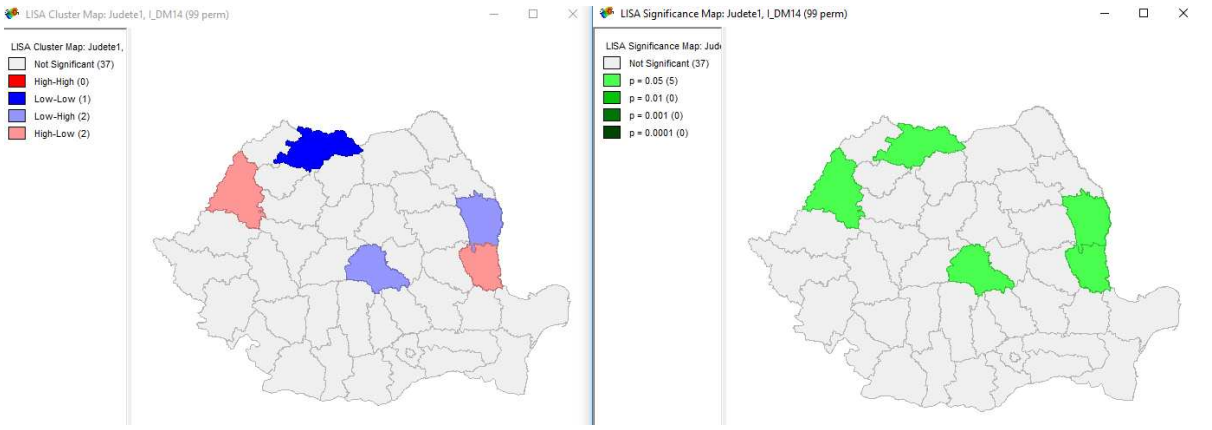
(a) (b)

Figure 4: The clusters based on CORT distance (a) and ACF distance (b)
Source: own processing

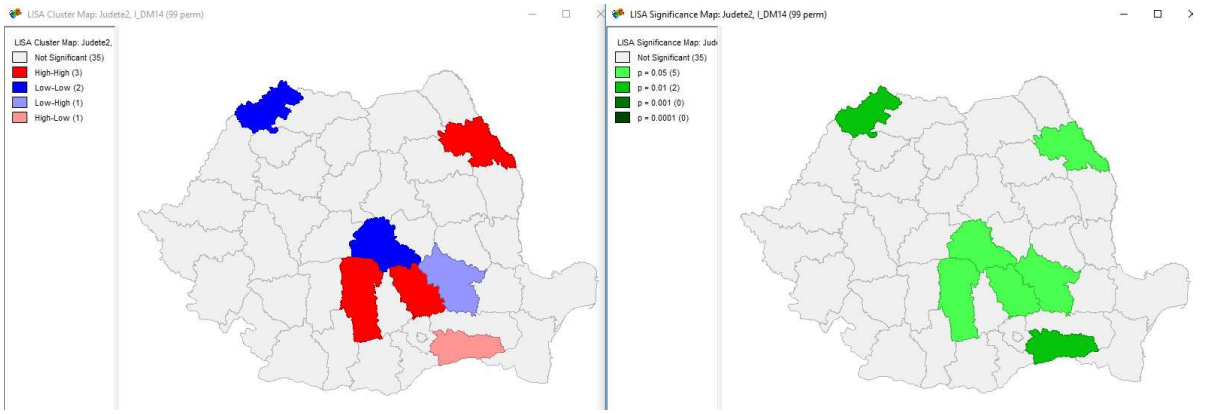
Figure 5 illustrates spatial patterns of malnutrition-related diabetes from the perspective of the concentration of similar values in space. The cluster maps point to significant cases of local spatial association by type of spatial correlation: bright red for the high-high association, bright blue for low-low, light blue for low-high, and light red for high-low. The high-high and low-low locations suggest clustering of similar values of malnutrition-related diabetes, whereas the high-low and low-high locations indicate spatial outliers. The associated significance maps identify the counties having significant local Moran statistics.

The results of LISA analysis in the form of a cluster map (i.e. a map with locally specific values which may form visual clusters) and a significance map answer

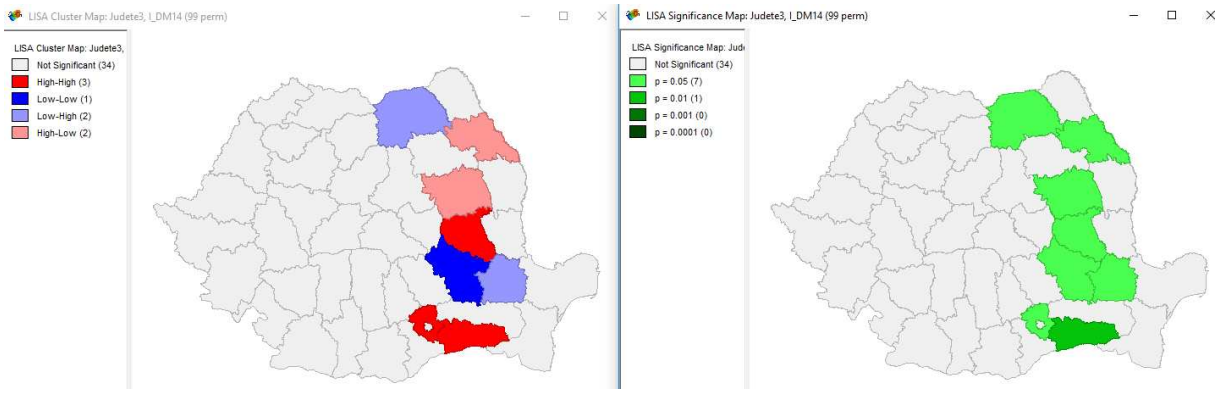
important questions, such as where the clusters can be found, what they look like, and whether they are random or statistically significant. The spatial weighting scheme defines which units are considered geographically close for the calculation of spatial autocorrelation.



(a)



(b)



(c)

Figure 5: LISA spatial clusters maps for malnutrition-related diabetes and the associated significance maps, for different weighting schemes: first-order contiguity (a), second-order contiguity (b) and third-order contiguity (c)

Source: own processing

The LISA spatial clusters depicted in the maps in Figure 5 reveal a variety of cases. The spatial weighting scheme chosen for the analysis of spatial dependence is to a great extent arbitrary (Nosek and Netrdová, 2014) although it significantly affects the final outcomes. We apply spatial local clustering for three weighting variants, using first, second and third order contiguity. Naturally, a higher order of contiguity means more local clusters because we allow for larger interactions in space.

If only first neighbors are considered, most clusters point to dissimilarities between direct neighbors, except for Satu-Mare and its surrounding counties: When the neighborhood is extended to include the second and third neighbor of a county, all four types of spatial clusters emerge on the maps.

The results from the model estimations (Table 5) reveal interesting differences between the two groups of models (with/without temporal cluster dummies) even in the significance of spatial interactions in malnutrition-related diabetes.

Table 5: The results from the regression models (dependent variable – – new cases of malnutrition-related diabetes)

Model 1 – without temporal cluster dummies						
Variables	**Classic model***		**Spatial lag model****		**Spatial error model****	
	Coeff.	**Prob.**	**Coeff.**	**Prob.**	**Coeff.**	**Prob.**
W_DM			-0.4707	0.0305		
Constant	-4.079	0.3178	-1.7138	0.0379	-2.3847	0.4872
GDP per capita	-0.0082	0.0782	-0.0077	0.0507	-0.0072	0.0708
Wage	0.0066	0.0924	0.0059	0.0808	0.0052	0.1139
MD(t-1)	0.5176	0.0002	0.5040	0.0000	0.4776	0.0000
LAMBDA					-0.3622	0.1369
Statistics	**Value**	**Prob**	**Value**	**Prob**	**Value**	**Prob**
R-squared	0.3876		0.4655		0.4276	
Log likelihood	-104.869		-102.923		-103.994	
F-statistic	8.018	0.0002				
Breusch-Pagan test	3.1336	0.3748	2.6537	0.4482	2.7036	0.4396
Koenker-Bassett test	1.8030	0.6127				
Spatial dependence: Likelihood Ratio Test			3.8930	0.0404	1.7512	0.1857

Model 2 – with temporal cluster dummies				
Variables	**CORT distance Classic model***		**ADF distance Spatial lag model****	
	Coeff.	**Prob.**	**Coeff.**	**Prob.**
W_DM			-0.4927	0.0280
Constant	6.4899	0.0911	-1.6244	0.6993
GDP per capita	-0.0076	0.0460	-0.0091	0.0520
Wage	0.0056	0.0806	0.0068	0.0851
Temporal cluster dummies (reference group – cluster 4)				
Cluster 1	-9.8028	0.0000	2.1886	0.0420
Cluster 2	-4.4290	0.0324	insignificant	
Cluster 3	-8.5525	0.0000	insignificant	
Statistics	**Value**	**Prob**	**Value**	**Prob**
R-squared	0.6202		0.2885	

F-statistic	11.7558	0.0000		
Log likelihood	-94.84		-109.015	
Breusch-Pagan test	11.2824	0.0491	4.1323	0.2475
Koenker-Bassett test	10.7175	0.0573		
Spatial dependence: Likelihood Ratio Test			3.6332	0.0466

**OLS estimation*

*** Maximum likelihood estimation*

In Table 5, the first group of models clearly shows the higher relevance of the spatial lag model against the alternatives – classic model and spatial error model, based mainly on the spatial dependence Likelihood Ratio Test, which indicates it as the best fit for our data. What could be the cause of this evident spatial dependence in the distribution of a disease that is not, from a medical perspective, driven by contagion effects? We believe that the answer could be the common underlying factors of influence that neighbor counties share. Since GDP per capita and average wage are already in the models, we have to consider other factors that determine the spatial autocorrelation in the emergence of new cases of malnutrition-related diabetes. As malnutrition-related diabetes is believed to be strongly influenced by diet, these underlying factors could be regional eating habits (including traditional local cuisine based on certain ingredients, recipes and cooking methods more prone to triggering diabetes), common regional life styles, e.g. preference for home-cooked food versus dinning out, the latter being a less healthy option, prevalence of fast-food in the diet of young people, etc.

The variable GDP per capita, used as proxy for the general development level of the counties, is significant and bears a negative sign in all models, therefore it indicates the negative impact of the lower health status of the population in poorer counties and the influence of the drawbacks in the health care system, a key factor in prevention, early diagnostic and treatment of illnesses.

Although wage is a factor of influence having low significance level in the models, its positive sign confirms that one should rule out poverty driven food deprivation as causal factor and focus on the more likely inappropriate eating habits, for instance less home-made food versus more convenient (albeit more expensive) commercial food.

Finally, the one year lag of the dependent variable is positive and highly significant in all models in the first group, suggesting short-term persistence, stability in the emergence of new cases of malnutrition-related diabetes. This variable is dropped from the second group of models since the temporal clusters convey the same information, especially when using the ACF distance that captures the time series memory from one year to another.

The second set of models, including the temporal cluster variables, largely confirm the previous findings regarding the influence of GDP per capita and wage, and it is noteworthy that both variables become more statistically significant in the new specifications. The models based on clusters emerged from the two variants of distance give entirely different results, which was expected given that they convey distinct information: the CORT distance captures similarity in the shape of the time series, while the ACF distance might be considered a measure of time series memory from one year to another. The CORT distance is by far more appropriate for our data since it provides a higher explanatory power (62.02% against 28.85%) and all dummy variables capturing the time series clusters are highly significant (Table 5, model 2). Cluster 4 has been selected as the reference group because it had the most irregular behavior among the four, alternating from periods with high incidence to periods with low incidence. Since it includes counties having many new cases of

malnutrition-related diabetes, the dummy variables for the other three clusters all bear negative signs.

An interesting outcome from the model based on CORT distance clustering is that spatial models had to be dropped for the first time in favor of the classic regression model. The most likely explanation is that the cluster variables are able to capture the spatial dependence as well. Although the temporal clusters are built based on common dynamics not contiguity (as is the case of spatial clusters), many counties belonging to the same temporal cluster are neighbors (see the time series cluster mapping in Figure 4). In this case, the spatial autocorrelation being included in the cluster regressors, the spatial models become useless.

In sum, standard statistical tests validate the spatial lag model in the first group and classic OLS regression in the second group of models as best specifications. It should also be noted that diagnostics for heteroskedasticity are much better for all models in group 1.

5. Conclusions

Based on a data set provided by CNSISP covering a time span between 2007 and 2014, and a GINI index calculation, we found that among the nine categories in diabetes incidence and prevalence the highest inequalities across Romania correspond to malnutrition diabetes. Therefore, in our paper, the focus of attention was on exploring novel ways of combining information from different but complementary types of clustering methods (temporal versus spatial) in order to gain a deeper understanding on the spatial distribution of this rather neglected disease.

There are three categories of contributions and findings that deserve to be emphasized. First and the foremost this paper is an illustration of a time series clustering application for real data. While the temporal clustering captures a lot of interest among specialists and the literature is reach in theoretical considerations, the applications are still scarce and even those that are available prove to be rather abstract in their interpretations.

Secondly, and as important as the previous contribution, we combined two techniques relatively new in Romanian research: spatial analysis and time series clustering. From a methodological perspective, our main finding was that the temporal and spatial analyses complement each other quite well and that the time series cluster variables are able to capture the spatial dependence as well. More precisely, we showed throughout our analysis that by changing the dissimilarity distance in the initial stage of temporal clustering, we can capture both spatial and temporal features of the data, as suggested by the statistical tests indicating that a spatial model is no longer necessary for capturing the spatial dependence present in our data.

Last, but not least, we discussed the malnutrition – related diabetes, mellitus, as a possible proxy of poverty, and tried to advocate our claim by relating this disease's distribution with some economic variables. At a first glance, the findings did not confirm entirely our assumptions. On the one hand, we found indeed a negative impact of the general development level of the counties (proxied by the variable GDP per capita). On the other hand, wage didn't prove a significant negative impact, as expected. In our opinion, this result is an invitation for further and more focused research: Romania is a middle level income country, and the malnutrition is definitely different in its essence than its counterpart in developing countries. What we expect in Romania might be rather related to "modern malnutrition" mentioned in an increasing body of literature, correlated with bad consumption habits and lack of relevant nutrients. We could speculate that our findings point toward a relation between wage and malnutrition – related diabetes somehow similar in its nature with the demand for inferior goods: as the salary increases, the demand for these goods decrease but only if a threshold in purchase power is reached. Below that threshold, the consumer will

be tempted to keep consuming inferior goods, and may even increase their consumption. If we accept a level of significance of 10% for our "wage" variable, this is a possible interpretation. Otherwise, we can admit that the variable is not statistically significant at 5% level, most likely pointing to the fact that the wage overcame a critical threshold and it is not anymore relevant to explain this type of malnutrition – related disease.

Another argument that can be taken into account when interpreting the results is that, according to the medical literature, it may take ten years or more for this condition to manifest after a severe nutrients deprivation. Unfortunately, relevant data are not available for such an investigation. In any case, and in relation to our very concrete result, drawing attention on life style and diet risk factors as determinants for malnutrition diabetes, and raising awareness on the importance of education and better information for preventing the emergence of new cases are both desirable and possible.

Potential further research derives from the unavoidable limits of our work. One of the most important limits come from the fact that we tried to explain incidence in a type of disease using data regarding both people affected and not affected by that condition. Only conducting a study on patients, and not using official data can help to overcome this limit. We used a limited number of variables: most likely the distribution of medical specialists across Romania, or other explanatory variables can be relevant as well.

References

Anselin, L. (2005), Exploring Spatial Data with GeoDaTM : A Workbook, Spatial Analysis Laboratory Department of Geography, University of Illinois, Urbana, available at: http://sal.agecon.uiuc.edu.

Anselin, L., Rey, S. (1991), "Properties of Tests for Spatial Dependence in Linear Regression Models*", Geographical Analysis*, 23, pp. 112–131.

Banerjee A., Dave, R.N (2004) Validate cluster tendency using Hopkins statistics, Conference Paper *in* IEEE International Conference on Fuzzy Systems 1:149 - 153 vol.1 / 54, August 2004

Blakely T, Hales S, Woodward A. (2004) Poverty : assessing the distribution of health risks by socioeconomic position at national and local levels. Geneva, World Health Organization, 2004. (WHO Environmental Burden of Disease Series, No. 10).

Chiriac C, Scorțan A. (2015) Evidența diabetului zaharat în perioada 2007 – 2014, Raport al Ministerului Sănătății, Institutul Național de Sănătate Publică, Centrul Național de Statistică și Informatică în Sănătate Publică

Fotheringham, A. S. (1997), "Trends in quantitative methods I: stressing the Local", Progress in Human Geography 21 (1), pp. 88-96.

Galeano P, Pena D (2000). "Multivariate Analysis in Vector Time Series." Resenhas do Instituto de Matematica e Estatıstica da Universidade de Sao Paulo, 4(4), 383–403.

GeoDa (2014), *The GeoDa Center for Geospatial Analysis and Computation*, available at: http://geodacenter.asu.edu/about

George AM, Jacob AG, Fogelfeld L. (2015) Lean diabetes mellitus: An emerging entity in the era of obesity. *World Journal of Diabetes*. 2015; 6(4):613-620. doi:10.4239/wjd.v6.i4.613.

Hotez P., Herricks J. (2015) Helminth elimination in the pursuit of Sustainable Development Goals: a "worm index" for human development (*PLOS Neglected Tropical Diseases*, 30 April 2015)

Kaufman L., Rousseeuw P. J. (2005) Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley

Kidd S., Wylde E., (2011) Targeting the poorest: An assessment of the proxy means test methodology, Published by the Australian Agency for International Development (AusAID), Canberra, September 2011, available at http://www.unicef.org/socialpolicy/files/targeting-poorest.pdf, retrieved on 8th November 2016

LeSage, J.P., Pace R.K. (2009), Introduction to Spatial Econometrics, Boca Raton, CRC Press

Lester F. T. (1993) A search for malnutrition-related diabetes mellitus among Ethiopian patients. Diabetes Care. 1993 Jan; 16(1): 187–192.

Liao T. W., (2005) Clustering of time series data – a survey, Pattern Recognition (38) (2005), 1857 – 1874

Montero P., Vilar J (2014) TSclust: An R Package for Time Series Clustering, Journal of Statistical Software, November 2014, Volume 62, Issue 1.

National Institute of Statistics: TEMPO database - time series, 2016, available at https://statistici.insse.ro/shop/

Nosek, V., Netrdová, P. (2014), Measuring Spatial Aspects of Variability. Comparing Spatial Autocorrelation with Regional Decomposition in International Unemployment Research, Historical Social Research, Vol. 39, 2, pp. 292-314.

Online reference: Food poverty and health, a contribution of the Faculty of Public Health of the Royal College of Physicians in the UK, available at http://www.fph.org.uk/uploads/bs_food_poverty.pdf retrieved on 9 November 2016

Paparrizos J., Gravano L., (2015) K-Shape: Efficient and Accurate Clustering of Time Series, Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, p. 1855-1870

Press V. (2004) Nutrition and food poverty: a toolkit for those involved in developing or implementing a local nutrition and food poverty strategy. London: National Heart Forum

Samuels F., Pose R. R. (2013) Why neglected tropical diseases matter in reducing poverty, Working Paper for Development Progress, UK, available at http://unitingtocombatntds.org/sites/default/files/resource_file/Why%20neglected%20tropical%20diseases%20matter%20in%20reducing%20poverty.pdf, retrieved 8 November 2016

Schneider M.C. et al. (2002) Trends in infant mortality inequalities in the Americas: 1955 – 1995, Journal of Epidemiology Community Health 2002; 56:538-541 doi:10.1136 / jech.56.7.538

Taksande A., Taksande B., Kumar A., Vilhekar KY (2008) Malnutrition-related Diabets Mellitus, JMGIMS, Vol. 13, No(ii),19-24

Unwin, A., Unwin, D. (1998), "Exploratory Spatial Data Analysis with Local Statistics", The Statistician, 47 (3), pp. 415-421.

World Health Organization. Diabetes mellitus. Tech Rep Ser 1985; 727: 20-4.

Worrall E., Basu S., Hanson K. (2003) The relationship between socio – economic status and malaria: a review of the literature, Background paper for "Ensuring that malaria control interventions reach the poor", London, 5 – 6 September 2002, available at http://siteresources.worldbank.org/INTMALARIA/Resources/SESMalariaBackgroundPaper.pdf, retrieved on 8th November 2016