DOI: 10.22616/j.balticsurveying.2018.018

POSSIBILITIES USE TO SELECTED METHODS OF SPATIAL DATA MINING IN DEMOGRAPHIC DATA ANALYTICS

Krystyna Kurowska, Ewa Kietlinska, Hubert Kryszk
Faculty of Geodesy, Geospatial and Civil Engineering
University of Warmia and Mazury in Olsztyn, Olsztyn, Poland

Abstract:
The main purpose of data mining in private and public sector institutions is to process and analyse data with the aim of generating reliable information for decision-making. Decision-making performance is determined by the availability of the relevant data and the user's ability to adapt that data for analytical purposes. The popularity of spatial statistical tools is on the rise owing to the complexity of the analysed factors, their variation over time and their correlations with the spatial structure.
Popular models should be applied in demographic analyses for the needs of the spatial planning process. The availability of high-resolution data and accurate analytical tools enhances the value of spatial analyses, and the described models can be universally applied to support the decision-making process.
The aim of this study was to present the applicability of selected spatial statistical models for analysing demographic data in the planning process and to identify the main advantages of these models.
Keywords: spatial data mining; spatial statistics; autocorrelation; demographic analysis.

Introduction

Cities play a key role in human settlement systems and they are the main hubs of economic activity. Depending on their rank in the settlement systems, cities are local, regional or national growth centres whose development influences the surrounding areas (Kowalczyk 2015). Large urban areas drive demographic changes and, consequently, changes in population structure. For this reason, the economic status of cities and the economic processes that take place in urban areas are of vital significance for the national economy (Lee and Rinner, 2015).
Demographic processes such as natural population growth and migration are directly responsible for changes in the size and structure of the urban population. According to the literature, the rate of these processes is influenced by changes in the standard of living and quality of life, growing levels of social mobility, the situation on the job market, income levels, the quality of health care services and educational attainment (cf. Cigno, 1991; Cliquet, 1991; Kotowska, 1999).
Purely demographic processes evolve at different rates and with different intensity, and they exert a growing influence on various areas of life, including the economy, the job market and the spatial management of territorial units. The above increases the demand for data analyses and forecasts based on various phenomena (Kurowska and Kietlińska, 2017).
According to Gaździcki (2001), information systems cater to the growing demand for the accumulation, processing and dissemination of data. Data that are in any way related to space are processed in spatial information systems which are defined as systems that process, accumulate, verify, integrate, analyse, transfer and disseminate spatial data.
Information systems provide users with a wide range of research tools, beginning with the simplest descriptive and statistical methods to complex mathematical models that support accurate analyses and forecasts of demographic processes and the associated economic processes.
Spatial data have a multidimensional structure. In addition to the basic information describing the location of an object in space (geographic coordinates), other types of data can also be taken into account during the analytical process, including variations in the observed phenomena over time and/or other qualitative or quantitative attributes describing the attributes of the analysed object. In view of the multidimensional character of this process and the complexity of the analysed data, statistical methods for describing, investigating and analysing object attributes should be developed based on classical statistical measures that have been suitably modified for this purpose (Sunmin et al., 2018).

In analyses of spatial databases, the investigated phenomena not only have to be quantified, but the mutual relations and interactions between the neighbouring objects also need to be described (Ramirez, Loboguerrero, 2002).
Spatial analyses can be performed on any type of objects localised in space. Spatial objects do not exist in isolation; therefore, their interactions with the surrounding space need to be investigated. According to Tobler's law (Tobler, 1970), objects that are separated by a smaller distance are characterised by stronger interrelations and more significant mutual interactions than distant objects.
Therefore, the definition of mutual interactions should be based on the definition of neighbourhood. In practice, this is accomplished with the use of spatial weights matrices to evaluate the impact of environmental factors on the processes investigated in a given region (Salamon, 2008).

Methodology of research and materials

The aim of the study was to analyse the applicability of the existing analytical models for analyses of demographic data. The results of such analyses provide highly valuable inputs for spatial planning and decision making. Special attention was paid to the availability of software and data for the proposed analyses.
There is a broad selection of commercial software and shareware programs for the visualisation and analysis of spatial data. Esri-ArcGIS, a popular mapping and analytics platform, was used in this study. The analytical packages in Esri-ArcGIS have different functionalities. Spatial Statistics tools are particularly useful for spatial analyses of demographic data. In this study, they were used to determine the dispersion of spatial phenomena and to perform multicriteria similarity analyses. Local and global statistics were described, and the results were visualised on thematic maps in the Esri-ArcGIS environment.
The applicability of demographic data in GIS programs was evaluated to determine whether the use of such solutions in this study was justified. The research was based on the case study of Warsaw – the biggest city and capital of Poland.
The selection of the relevant data is one of the greatest challenges facing the user before analysis. Data relating to the units of administrative division (Warsaw districts), census districts and address points were selected depending on the type and scale of the conducted research and the size of the analysed area.

Discussions and results

The applicability of spatial data mining methods for analyses of demographic data

Information is one of the most valuable commodities in the modern world. Computer users are bombarded with vast amounts of electronic data on a daily basis. The potential benefits of the accumulated data are determined by the user's ability to process that data and develop reports, identify similarities or trends. Skilful decision-making based on the results of statistical analyses, inference and data use supports the achievement of business, operational and scientific goals.

Knowledge Discovery in Databases, an emerging field of research, supports the extraction of useful knowledge from the rapidly growing volumes of data. KDD is closely related to statistics, in particular data mining methods. According to Gregory Piatetsky-Shapiro (1995, 1996, 2007), the co-founder of the KDD, knowledge discovery is "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data".
Data mining is a stage of knowledge discovery. Large datasets are analysed with the use of selected algorithms to discover data patterns. A data model is developed, and its ability to accumulate useful knowledge is evaluated during the interactive KDD process which usually relies on subjective human judgement.
Spatial data mining is a specific process due to the unique character of the analysed data. Spatial data are often characterised by complex interactions and non-homogeneity, and observations that are separated by a small distance in geographic space are mutually related (are similar or different, depending on the observed phenomenon). Many geographic processes have a local character, and they are characterised by spatial non-homogeneity and instability relative to location.

Various models can be used to capture the above phenomenon and analyse spatial data. Standard deviation is the basic statistical measure for analysing clusters of the observed data. However, spatial units are not distributed uniformly in all directions from the central tendency; therefore, classical distribution analyses do not fully reflect the nature or dimensionality of the described objects. This problem can be resolved by calculating the standard distance separately for every direction to accurately determine the distribution of objects in space. The resulting values denote the axes of a standard deviational ellipse whose shape and direction reflect the orientation of object distribution. The axes of symmetry of the ellipse are rotated by a given angle to reveal the direction of dispersion around the central tendency and the direction and extent of minimum and maximum dispersion (Suchecka, 2014).

The directions of dispersion and object clusters can be examined globally for the entire set of points, or in intervals based on the spatial variability of objects in time. In demographic and behavioural studies, this tool is highly useful for analysing changes in population distribution. It can be used to describe the degree of clustering and dispersion, but when time is factored in as an additional factor, the discussed tool supports a comprehensive description of the analysed phenomena, which is of utmost importance in spatial planning.

In addition to analyses of data dispersion and data variation over time, GIS tools can also be used for more detailed and sophisticated analyses of demographic data. However, this process is fraught with many difficulties due to the complex nature of demographic data as well as the influence of social, environmental, technical and spatial factors. Such analyses should involve the largest possible volume of data, both quantitative and qualitative.

Demographic analyses involve numerous factors, and GIS tools seem to be well suited for comprehensive analyses of spatial data with the aim of identifying the searched attributes.

However, despite the wide availability of modern tools and complex analytical techniques, users often have to choose between several alternatives when making spatial decisions. The main aim of multicriteria analyses is to select the optimal variant, which incorporates difficult to compare criteria that significantly influence a given solution.

In demographic and spatial studies of population distribution, multicriteria analyses can be used to plan the location of objects (schools, healthcare facilities), administer territorial units and plan transportation routes.

Suitability (similarity) maps are an interesting solution in multicriteria analyses. This tool supports the identification and ranking of similarities between objects. In analyses that rely on a reference object, objects whose parameters are most similar to the reference parameters can be selected from a set of potential objects. A reference object can be an individual object as well as a group of objects with the desired attributes.

Similarity is determined based on a set of selected attributes with the use of one of three methods. Objects can be classified based on attribute values, the sequence of attributes in a series, or the relationships between variables (ESRI, 2010). In the classical approach, the attributes of the candidate objects are merely compared with the parameters of the reference object. The attributes are always determined by the type of analysis, and when standardised appropriately, they can be used in evaluations of similarity. The applicability of a conventional similarity analysis can be expanded by developing a continuous map of the investigated area. The map is used to identify objects that are most similar to the reference object in terms of the specified attributes.

Objects can be analysed not only for the presence of similarities, but also mutual relations. There are two measures of spatial autocorrelation in spatial statistics: global measures and local measures. Global autocorrelation denotes the presence of spatial relations between variables in the entire unit, whereas local measures indicate spatial relations between the variables in a given location and the variables in the neighbouring locations (Ord and Getis, 1995, 2011).

In line with the above definition, global spatial autocorrelation supports the determination of the mutual relations between objects in the entire unit, and it is usually investigated with the use of global Moran's I statistic (Siano, D'Uva, 2011). Moran's I has a positive value when the studied objects are similar and a negative value when similarities are not found. The value of Moran's I approximates zero when objects are randomly distributed (no autocorrelation). The main disadvantage of global autocorrelation is that its values are determined by the aggregation of the entire dataset into regions.

Both global and local attributes of the evaluated dataset are taken into account in detailed spatial analyses. The correlations in the entire studied unit are determined in a global analysis, but the significance of an individual object in its immediate neighbourhood can be determined only in a local analysis.

Local Indicators of Spatial Association (LISA) statistics are the most popular and widely used measures of local autocorrelation. These indicators are applied to determine the similarity of a spatial unit relative to its neighbours and the statistical significance of the observed relationship (Anselin, 1995). Local indicators of spatial association include local Moran's I, which supports the determination of the spatial effects of agglomeration, and local Geary's C which reflects spatial similarities and differences (Geary, 1995). Moran's I is applied to determine whether the investigated area neighbours regions with similar values of the studied variable relative to the random distribution of these values in space. Similarly to global statistics, local Moran's I has a negative value if the neighbouring area differs significantly from the analysed area. Positive values of Moran's I denote similarities between the investigated region and its surroundings.

A demographic analysis of the Warsaw area

The rate of population growth in Warsaw was determined by analysing changes in the population of Warsaw districts between 1950 and 2012. An analysis of changes in the observed trend over time supports preliminary data mining, identification of data patterns and, consequently, in-depth data analysis.

The results of the analysis indicate that until 1970, the local population was concentrated mainly in the pre-war districts of Warsaw (Mokotów, Ochota, Wola) that had been reconstructed after the war. Beginning in the early 1970s, the construction of high-rise residential districts contributed to the gradual spread of the local population to other parts of the city.

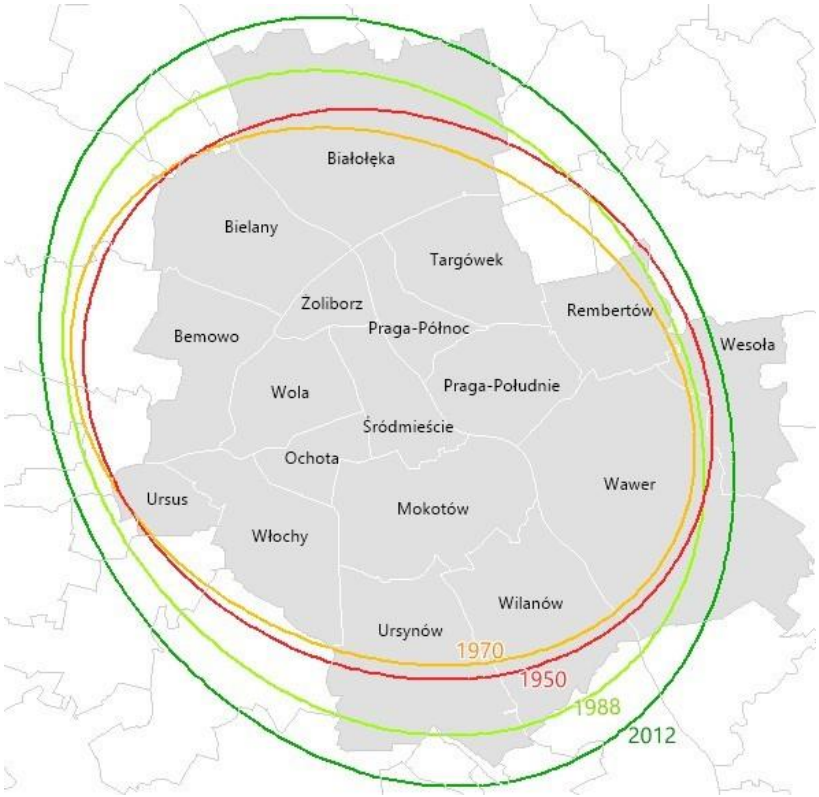

Fig. 1. Concentration of Warsaw's population between 1950 and 2012 (Kurowska and Kietlińska, 2017).

The political transformations of 1989 initiated profound economic and demographic changes in Poland. The availability of mortgage loans, rapid economic growth and rising incomes in urban areas contributed to suburbanization. The liberalization of the real estate market and the reinstatement of

land rents were also important drivers of urban to suburban migration (Słodczyk, 2001). Suburbanization produced a negative net migration rate in large cities and a positive net migration rate in the surrounding areas. Research indicates that not all metropolitan municipalities develop at the same rate. The highest population growth and the highest migration rates are noted in the municipalities adjacent to the urban core, whereas the urban population increases at a modest rate (Kasanko et al., 2006; Pardo-García and Mérida-Rodríguez, 2018; Sroka et al., 2018). Uncontrolled migration to suburbia changes the function and appearance of those areas (Kajdanek, 2011). The concentration ratio of Warsaw's population in the past 60 years has been calculated by Gawryszewski (2010).

Table 1

Lorenz concentration ratio for 18 districts in Warsaw in 1950 – 2010

	Year						
	1950	1960	1970	1978	1988	2002	2010
Lorenz concentration ratio ($0 \leq k \leq 1$)	0.590	0.605	0.618	0.589	0.531	0.474	0.438

Cities undergo constant change. Their architecture, infrastructure, public spaces and the quality of human resources continue to evolve over time. The availability of services, including public transport and culture, significantly contributes to a city's appeal. Good living conditions improve living standards in a city and the entire metropolitan area. However, not all urban residents have equal access to infrastructure, services and public spaces. In many cities, social polarisation leads to spatial segregation. These problems often stem from inadequate spatial development policies which focus on improving living standards in downtown areas, environmentally-friendly housing estates, business parks and shopping centres, but disregard areas that are less attractive for investors.

For this reason, demographic analyses should not only investigate the distribution of populations, but also the spatial distribution of different social groups. Various tools can be used for this purpose.

Similarity analyses are highly useful for in-depth evaluations of urban demographics. They support the identification of population groups characterised by various attributes. Given the nature of the analysis and the type and detail of the available data, the incomes of Warsaw residents were mapped on the assumption that wealth can be evaluated based on two factors: the type of residential building and per-capita buying power.

For continuous visualisation of the analysed data, address points were aggregated in a 250x250 m grid to evaluate the degree of similarity in the studied area. This approach was adopted to analyse microdata and to avoid errors resulting from excessive data aggregation. Per-capita incomes were expressed in Polish zloty (PLN) in a grid cell. The proportion of the population inhabiting various types of buildings (single-family or multi-family housing) was expressed as a percentage. The reference point (grid cell) was a point characterised by the highest value of the per-capita buying power index (PLN 22896/year/person) where 100% of the population inhabit single-family houses. Uninhabited grid cells were excluded from the analysis.

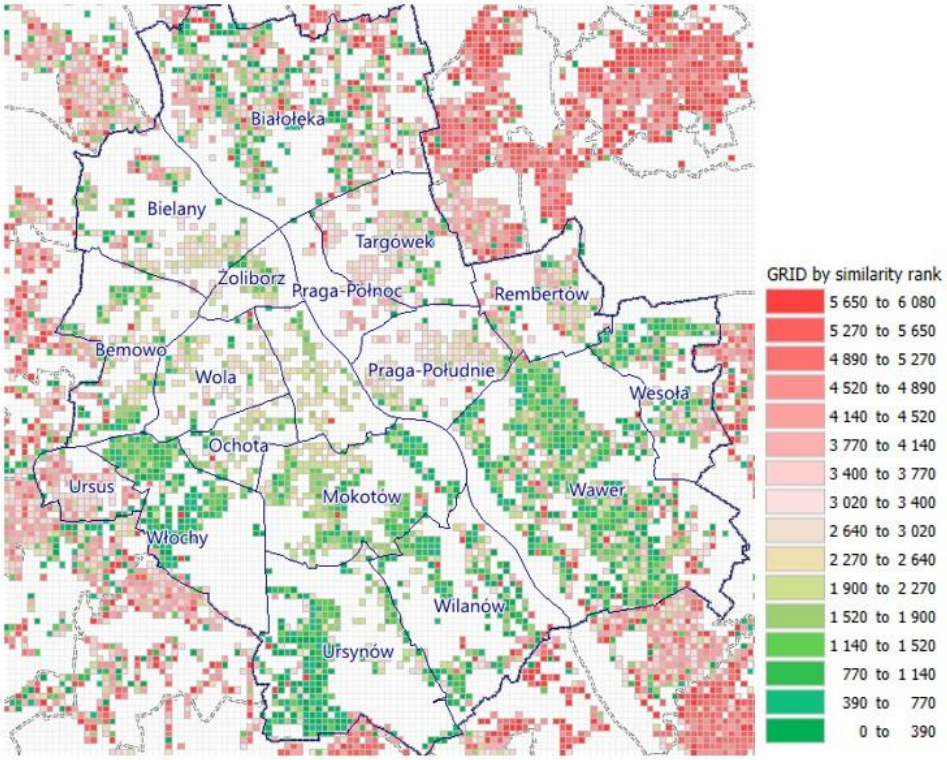

Fig. 2. A map of similarities in the incomes of Warsaw residents (data for 2014).

All observations in the dataset (grid cells) were ordered based on their similarity to the reference point and were assigned a similarity index (the closer the examined value to the reference point, the lower the value of the similarity index). The resulting values were used to map incomes in different districts of Warsaw. South Warsaw was characterised by the highest levels of affluence (green grid cells).

The presented tool is highly useful for analysing demographic data. Only two values were analysed in this study (income, type of residential building), but the developed tool is highly versatile and it supports the implementation of diverse variables. Demographic data can be used as the main variable or as one of many variables in a complex analysis. When the extent of data aggregation and the range of attributes (income, age group, working age) are selected accordingly, the proposed tool can be used to identify target groups, including outside the context of spatial management and planning.

The demographic structure not only reflects income levels, but also the distribution of age groups in a population. Residents belonging to different age groups form distinct clusters in urban areas. These areas are strongly correlated and mutually related. Local and global statistical data were used and districts with the highest proportion of the working-age population inhabiting single-family homes were selected to determine the distribution of various age groups in Warsaw. The adopted method can be used to identify areas with a predominance of target groups characterised by selected attributes.

The choice of the optimal reference unit is a very important consideration. The proposed aggregate largely determines the extent to which the examined phenomenon can be accurately described. Artificially generated boundaries do not fully reflect the character of the analysed area or the existing limitations to human activity.

Census districts were used in this study. These units are generated artificially, but their main advantage over automatically generated statistical grids is that the shape of every census district is highly dependent on topography, the distribution of transportation routes and other obstacles (the human tendency to form groups was taken into account). A census district is defined as a spatial unit, which is created for the needs of a census or statistical research. To ensure that census operations are conducted efficiently, the size of a census district should not exceed 500 persons and 200 apartments. The analysis relied on demographic data for 2014.

In view of the above, it can be assumed that every census district has similar demographic potential regardless of its area. However, Warsaw is a highly urbanised area and, according to Central

Statistical Office requirements, census districts have to incorporate entire buildings (regardless of the number of apartments in a building); therefore, the criteria associated with the number of inhabitants and the number of apartments are often exceeded. For this reason, the proportion of the working-age population in the total population of a census district was used as a variable in the analysis.
The absolute value of single-family homes also varies in census districts and does not constitute a conclusive point of reference; therefore, the percentage of residential units in single-family homes in the total number of residential units in a census district was calculated for the needs of the analysis.
Spatial autocorrelation in census districts was analysed with the use of a queen contiguity-based weights matrix. These spatial weights were selected because they are completely independent of the size of census districts and because they focus on and significantly influence units in the immediate vicinity of the analysed object.
The results of spatial autocorrelation analyses involving Getis-Ord G and Moran's I statistics revealed the presence of strong spatial autocorrelations. The p-value was zero in both cases; therefore, the probability that objects were randomly distributed in space was negligible.

Table 2

The results of a spatial autocorrelation analysis

		Single_perc	**Work_perc**
Global Moran's I	Moran's I	0.735642	0.261534
	z-score:	154.514681	54.959793
	p-value:	0.000000	0.000000
Getis-Ord General G	Observed General G:	0.001014	0.000435
	z-score:	113.493413	5.976728
	p-value:	0.000000	0.000000

The maps created based on the Getis-Ord statistic present the distribution of clusters with high and low values of the analysed variables in Warsaw. The maps reveal distinct clusters of high values: a predominance of the working-age population in the districts of Tarchomin, Bemowo and Wilanów, and a predominance of single-family homes in the peripheral districts and suburban areas. Areas with a low proportion of single-family homes (cold spots) are found mainly in central Warsaw and, to a much lesser degree, in the neighbouring urban areas.

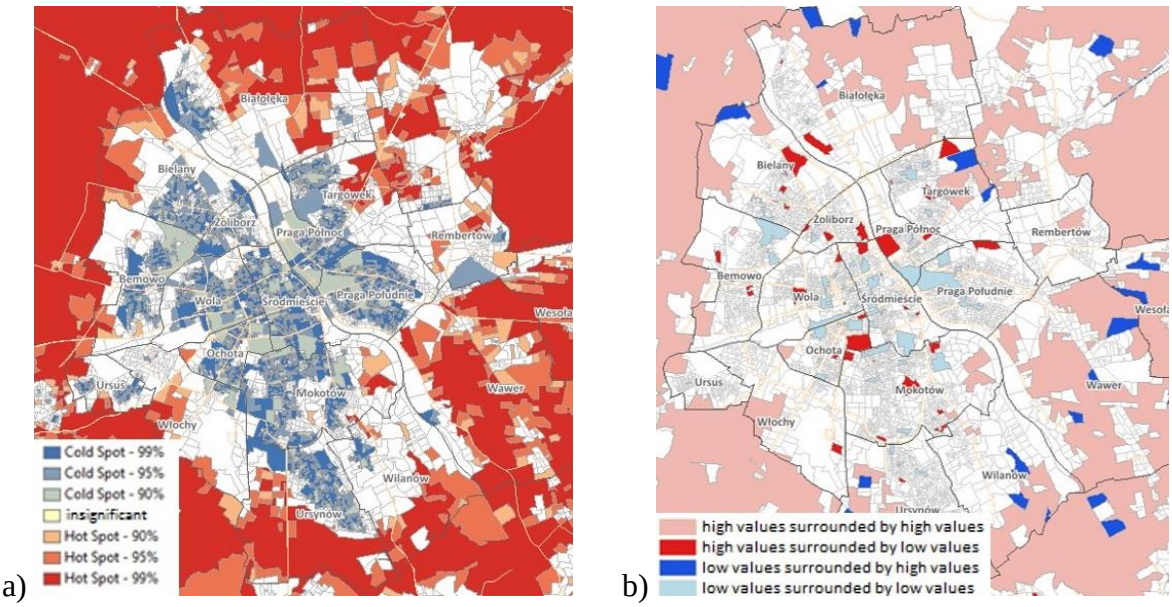

Fig. 3. Visualisation of clusters with high and low proportions of single-family homes: a) global correlation coefficients, b) local correlation coefficients (own elaboration).

The high values of global Moran's I were confirmed by LISA statistics. A queen contiguity-based weights matrix was also used in an analysis of local statistics. A global analysis supports a general overview of the data and facilitates preliminary statistical analyses, but spatial structure and mutual relations can be evaluated in detail only in a local analysis. Local correlation coefficients were mapped to reveal the presence of individual units with a predominance of single-family homes in areas where clusters of multi-family buildings were identified in the global analysis.

a) b)

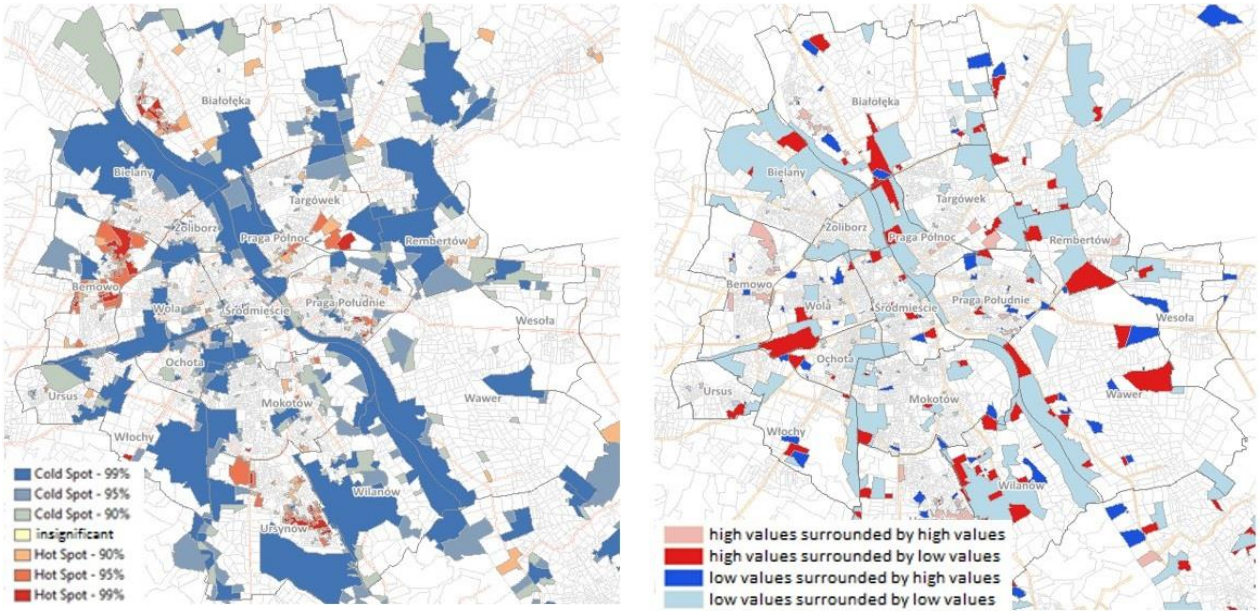

Fig. 4. Spatial distribution of the working-age population: a) global correlation coefficients, b) local correlation coefficients (own elaboration).

An analysis of global and local correlation coefficients relating to the working-age population in districts produced similar observations. A cohesive and homogeneous cluster of high values was identified in the district of Wilanów in the global analysis, but the presence of outliers where the values of the investigated variables were clearly lower than in the neighbouring districts was noted only in the local analysis.

Local statistics not only support the identification of significant clusters with similar values in the vicinity of the analysed unit, but they also provide valuable information about the spatial distribution and homogeneity of the analysed variable in a given area. Local statistics are also helpful in identifying atypical observations and clusters of high and low values.

However, the analysed phenomena are not always easy to describe with the use of two independent statistics. Districts with a predominance of the working-age population and single-family homes may be difficult to validate in this approach. In analyses that involve a higher number of variables, all attributes should be analysed jointly to select districts where the investigated variables reach maximum values.

Cluster analysis involves the search for natural clusters in datasets and data classification. Objects are grouped based on user-defined variables to minimise the differences between objects belonging to the same group and maximise the differences between objects belonging to different groups. Similarities are determined in the groups specified by the user or estimated by the tool.

The coefficient R^2 was calculated for every variable to determine the extent to which the variability of the original dataset was preserved during the grouping process. The accuracy of clustering increases with a rise in the value of R^2. The minimum, maximum and median values, standard deviation and the spread of the variables in the dataset and in groups are presented in Table 3.

Table 3

Statistics for the entire dataset and datasets generated by clustering

	Mean	**SD**	**Min**	**Max**	**R2**
single-family homes	0.3084	0.4095	0	1	0.9059
working-age population	0.6025	0.1606	0	1	0.7927
	Mean	**SD**	**Min**	**Max**	**Share**
single-family homes	0.0305	0.0845	0	0.449	0.449
working-age population	0.6363	0.0808	0.3333	1	0.6667
single-family homes	0.8671	0.1443	0.45	1	0.55
working-age population	0.6372	0.0625	0.25	1	0.75
single-family homes	0.1119	0.2916	0	1	1
working-age population	0.0022	0.0221	0	0.2778	0.2778

The most interesting results were obtained when the observations were divided into three independent groups (Table 3 and Fig. 5):

- blue – predominance of single-family homes and an above-average proportion of the working-age population;
- red – negligible proportion of single-family homes and an above-average proportion of the working-age population;
- green – negligible proportion of single-family homes and the working-age population.

Despite the fact that mutual spatial relations were not taken into account in the process of defining clustering parameters, clusters of objects belonging to different groups can be clearly identified in the map. This observation confirms the spatial dependence of the analysed variables.

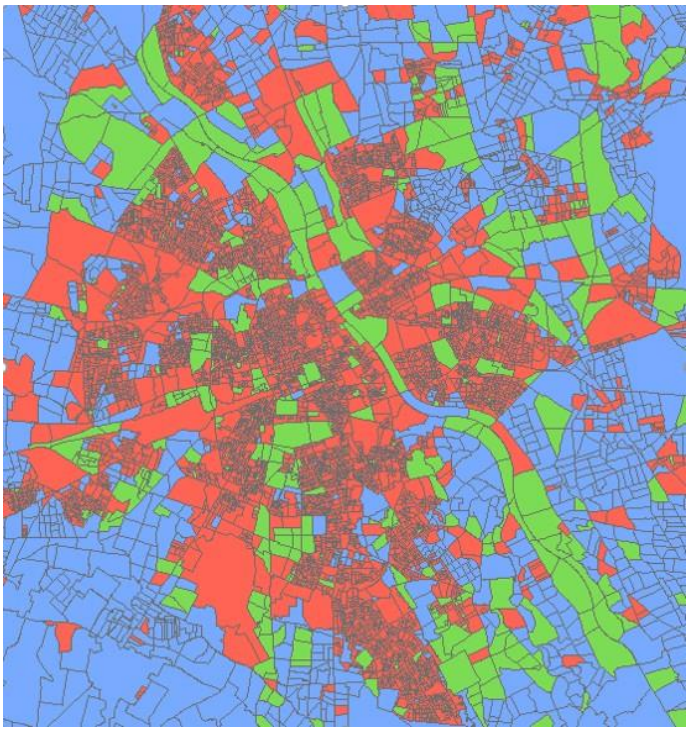

Fig. 5. Division of the urban space into groups based on the proportion of the working-age population and the proportion of single-family homes

The greatest benefits stemming from the implementation of local and global spatial statistics in traditional analytical processes include the identification of statistically significant clusters of high and low values, detailed analyses of the spatial distribution of the examined variable in a given area and the identification of outliers and atypical values.
In practice, the discussed solutions are applied by businesses and institutions to investigate the heterogeneity of a variable in the studied area. The visualisation of local statistics supports a quick identification of areas with low variable values, which cannot be achieved with the use of traditional analytical tools.

Conclusions

1. Cities play a key role in human settlement systems and they are the main hubs of economic activity. Depending on their rank in the settlement systems, cities are local, regional or national growth centres whose development influences the surrounding areas. The economic status of cities and the demographic processes that take place in urban areas are of vital significance for the national economy.
2. The growing applicability of software tools supports the broad use of methods and models exploring the location of spatial objects. The unique features of spatial data have to be taken into account in every stage of the process, beginning from the selection of data and the appropriate analytical methods to data visualisation. The user is responsible for the selection of the methods and tools that are best suited for analysing specific types of data.
3. mining requires methods that support detailed analyses of the examined variables and the identification of the type and degree of spatial autocorrelation, heterogeneity and interdependence. The choice of the appropriate visualisation techniques facilitates analyses of the spatial distribution of variables, determination of atypical locations and observations, and the identification of data patterns, clusters and special objects.
4. The choice of visualisation technique is determined by the scale of measurement, the number of dimensions and the user's preferences. Data transformations and calculations produce valuable spatial information for decision making.
5. Geographic information systems are valuable tools for mining demographic data. They can be used for simple visualisations of data as well as for planning, forecasting and modelling demographic processes. Detailed information about the location of spatial objects not only improves analytical accuracy but is an essential component of spatial research. It facilitates parameter prediction in any time interval, effective analyses of current data, demographic forecasts and evaluations of the influence of spatial correlations on strategic decision-making in a wide range of industrial applications.
6. The cross-sectional studies and reports generated with the involvement of GIS tools support comprehensive assessments of the observed phenomena and contribute to optimal decision making.

References

1. Anselin L. (1995) Local Indicators of Spatial Association. *Geographical Analysis*, 27, No. 2. p. 93-115.
2. Cigno A. (1991) *Economics of the family.* Oxford University Press, New York.
3. Cliquet R.L. (1991) *The second demographic transition: fact or fiction?* Council of Europe, Strasbourg.
4. Esri, (2010) ArcGIS Desktop 10 Help.
5. Gawryszewski A. (2010) Demographic and social development of Warsaw in the twentieth century. Mazovia Province. *Regional Studies*, 5, p. 11-28.
6. Gaździcki J. (2001) *Leksykon geomatyczny. [Lexicon of geomatics]*, Polskie Towarzystwo Informacji Przestrzennej, Warszawa.
7. Geary R.C. (1954) *The contiguity ratio and statistical mapping*. Incorporated Statistician, p. 115–145.
8. Kajdanek K. (2011) *Pomiędzy miastem a wsią. Suburbanizacja na przykładzie osiedli podmiejskich Wrocławia*. Nomos, Kraków.
9. Kasanko, M., Barredo, J. I., Lavalle, C., McCormick, N., Demicheli, L., Sagris, V., Brezger, A. (2006) Are European cities becoming dispersed?: A comparative analysis of 15 European urban areas. *Landscape and urban planning*, 77(1-2), p. 111-130.

10. Kotowska I.E. red. (1999) *Przemiany demograficzne w Polsce w latach 90. w świetle koncepcji drugiego przejścia demograficznego.* Wydawnictwo SGH, Warszawa.
11. Kowalczyk C. (2015) Dynamics of urban space. *Journal of Economics and Management* 19(1), p. 5-15.
12. Kurowska K., Kietlińska E. (2017) The applicability of accessibility analyses in spatial planning. *Baltic Surveying*, 49, p. 47-56.
13. Lee A.C.D., Rinner C. (2015) Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006. *Habitat International*, 45, p. 92-98.
14. Ord J.K., Getis A. (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27, p. 286-305.
15. Ord J.K., Getis A. (2011) Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation. *Journal of Regional Science*, 41, p. 411-432.
16. Pardo-García S., Mérida-Rodríguez M. (2018) Physical location factors of metropolitan and rural sprawl: Geostatistical analysis of three Mediterranean areas in Southern Spain. *Cities*, 79, p. 178-186.
17. Piatetsky-Shapiro G. (1995) Knowledge discovery in personal data vs. privacy: A mini-symposium. IEEE Expert: Intelligent Systems and Their Applications, 10(2), p. 46-47.
18. Piatetsky-Shapiro G. (1996) From Data Mining to Knowledge Discovery: an Overview. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, in *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.
19. Piatetsky-Shapiro G. (2007) Data Mining and Knowledge Discovery – 1996 to 2005: Overcoming the Hype and moving from "University" to "Business" and "Analytics", Gregory Piatetsky-Shapiro, *Data Mining and Knowledge Discovery Journal*.
20. Ramirez M.T. , Loboguerrero A.M. (2002) *Spatial dependence and economic growth: Evidence from a panel of countries.*
21. Salamon J. (2008) Studies on spatial autocorrelation of technical rural infrastructure development in the Świętokrzyskie province using Moran's I statistics. *Infrastructure and Ecology of Rural Areas*, 8, p. 207-214.
22. Siano R. De., D'Uva M. (2011) Italian regional specialisation: a spatial analysis. *Università degli Studi di Napoli-Parthenope, Discussion Paper*, 07.
23. Słodczyk J. (2001) *Przestrzeń miasta i jej przeobrażenia*. Wydaw. Uniwersytetu Opolskiego, Opole.
24. Sroka W., Mikolajczyk J., Wojewodzic T., Kwoczynska B. (2018) Agricultural Land vs. Urbanisation in Chosen Polish Metropolitan Areas: A Spatial Analysis Based on Regression Trees. Sustainability, 10(3), 837p.
25. Suchecka J. (2014) *Statystyka przestrzenna. Metody analizy struktur przestrzennych*. *[Spatial statistics. Methods of spatial structures analysis]* C.H.Beck, 2014.
26. Sunmin L., Saro L., Moung-Jin L., Hyung-Sup J. (2018). Spatial Assessment of Urban Flood Susceptibility Using Data Mining and Geographic Information System (GIS) Tools. Sustainability 10 (648); doi:10.3390/su10030648
27. Tobler W.R. (1970) A computer model simulating urban growth in the Detroit region. *Economic geography*, 46(2).

Information about authors
Krystyna Kurowska, PhD., Faculty of Geodesy, Geospatial and Civil Engineering, University of Warmia and Mazury in Olsztyn. Prawocheńskiego 15, 10-720 Olsztyn, phone: +48 895 234281, krystyna.kurowska@uwm.edu.pl Fields of interest: spatial planning, rural development, GIS.
Ewa Kietlińska, M.Sc., Faculty of Geodesy, SGN Sp. z o.o. Wyspiańskiego 14, 39-400 Tarnobrzeg, phone: +48 600 287 133, evakietlinska@gmail.com . Fields of interest: GIS, geostatistics, geoinformatics.
Hubert Kryszk, PhD., Faculty of Geodesy, Geospatial and Civil Engineering, University of Warman and Mazury in Olsztyn. Prawocheńskiego 15, 10-720 Olsztyn, phone: +48 895 234209, hubert.kryszk@uwm.edu.pl Fields of interest: real estate market, rural development, geoinformatics.