GEODESY AND CARTOGRAPHY
Vol. 61, No 1, 2012, pp. 47-70

DOI No 10.2478/v10277-012-0021-x

The statistical geoportal and the "cartographic added value" – creation of the spatial knowledge infrastructure

Anna Fiedukowicz[1], Jędrzej Gąsiorowski[2], Paweł Kowalski[1], Robert Olszewski[1], Agata Pillich-Kolipińska[1]

[1]Warsaw University of Technology, Faculty of Geodesy
and Cartography, Department of Cartography
Plac Politechniki 1, 00-661 Warsaw, Poland

[2]Institute of Geodesy and Cartography,
Department of Spatial Information Systems and Cadastre
27 Modzelewskiego St., 02-679 Warsaw, Poland
e-mail: r.olszewski@gik.pw.edu.pl; p.kowalski@gik.pw.edu.pl;
a.pillich@gik.pw.edu.pl; jedrzej.gasiorowski@igik.edu.pl

Received: 31 May 2012/Accepted: 15 July 2012

Abstract: The wide access to source data, published by numerous websites, results in situation, when ***information*** acquisition is not a problem any more. The real problem is how to transform information in the useful ***knowledge***. Cartographic method of research, dealing with spatial data, has been serving this purpose for many years. Nowadays, it allows conducting analyses at the high complexity level, thanks to the intense development in IT technologies, The vast majority of analytic methods utilizing the so-called *data mining* and *data enrichment* techniques, however, concerns non-spatial data. According to the Authors, utilizing those techniques in spatial data analysis (including analysis based on statistical data with spatial reference), would allow the evolution of the **Spatial Information Infrastructure** (SII) into the **Spatial Knowledge Infrastructure** (SKI). The SKI development would benefit from the existence of statistical geoportal. Its proposed functionality, consisting of data analysis as well as visualization, is outlined in the article. The examples of geostatistical analyses (ANOVA and the regression model considering the spatial neighborhood), possible to implement in such portal and allowing to produce the "cartographic added value", are also presented here.

Keywords: geostatistical analysis, geostatistical atlas, geoportal, spatial knowledge infrastructure

1. Introduction

In order to meet the demands of the modern information society, analyzing the space is equally important as its cartographically correct modeling. At present, the term "modeling" is understood, first of all, as creation of reference spatial data bases, which describe particular components of the natural environment and relations between them. Its second meaning is development of sectoral GIS systems, basing on the topographic

canvas, defined that way. Thematic data, stored in such systems, contains information of economic, demographic, social conditions etc. The wide access to source data, published by numerous websites, results in situation, when ***information*** acquisition is not a problem any more. The real problem is how to transform information in the useful ***knowledge***.

The so-called, cartographic method of research (CMR) has been known for more than a half of a century (Saliszczew, 1955, 1968; Berlant, 1978, 1986). This method consists of utilizing maps (and nowadays also geographical data bases) for description, analysis and scientific recognition of phenomena, discovery new regularities of their locations and mutual relations as well as forecasting changes. The essence of the CMR relies on inclusion of an intermediate tool, a map – as a model of researched phenomena – into the process of investigating the reality. Nowadays it is usually a digital map. The contemporary analytical CMR systems, using IT technologies, allow conducting analyses at the high complexity level (Poole et al. 1998). One of the ways of such analysis is to utilize the, so-called, *data mining* and *data enrichment*, i.e. to detect regularities, rules and structures "hidden" in the data base. The concept which is wider than data mining, is *knowledge discovery from databases* – KDD. It is based on the assumption that information is "hidden" in the data base in the form of the data *structure*. Data mining is often equalized with KDD; however, it is only one of the stages of knowledge discovery (Miller and Han, 2001). Data mining is the process of „data analysis which leads to getting information", while KDD is the „higher order process, which comprises analysis of information acquired by data mining and transforming it into *knowledge"*.

The vast majority of contemporary research works, which utilize data mining techniques, concern analysis of non-spatial data, i.e. simple tables which contain, for example, commercial information. Analysts detect structures and relations hidden in big files of data and they state, for example, that "customers, who buy the product A often also decide to buy products B or C". Research performed by the authors enlarges the scope of applicability of that method to analysis of spatial data, stored in geographical data bases. Utilization of data mining techniques allows researching large data resources (such as digital maps related to economic, demographic issues) in order to find patterns and interrelations (Wachowicz, 2001). Detection of rules "hidden" in the geographical data base, allows construction of a decision system, which, for example, enables to forecast trends of development of particular agglomerations or regions.

Therefore, data processing, including spatial data processing, serves for ***explicite*** presentation of relations, which ***implicite*** occur in the source data base (Olszewski, 2006; 2008). *Spatial data mining* techniques, often explained as the explorative analysis of spatial data, are the indispensable element of the "*data enrichment*" process. Presentation of spatial structures proves the power of the cartographic method of research. Analysis of spatial information, understood in this way, is therefore the source of creation of the "cartographic added value" – the knowledge which explains source data, its spatial arrangements and which allows forecasting trends of development. That

concept is fully compliant with the idea of the geoinformation structure development, proposed by Iwaniak (2011). He states, that after the period of development of the web **Spatial Data Infrastructure** (SDI) in the nineties of the 20^{th} century, allowing for locating, publishing and getting the access to spatial data, the beginning of the new century was connected with intensive development of the **Spatial Information Infrastructure** (SII), which allowed combining services, harmonizing and complex data processing. This also allowed for widening data publication over the web, with developed cartographic methods. However, following Iwaniak (2011), the further step towards the development of the geoinformation infrastructure is of key importance. Orchestration of spatial services, which has been recently observed, successive implementation of ontology, thesauruses and context processing of data published in the web, allowed evolution of the SII into the **Spatial Knowledge Infrastructure** (SKI). As a result of utilization of sophisticated algorithms of analyses and access to metadata, the infrastructure, understood in this way, is capable to "understand" data and services supplied by institutions, which distribute spatial data.

Following the authors' opinions, conditions of development of the SKI, defined by Iwaniak (2011) should be considered as necessary, but not sufficient. It is necessary not only to simultaneously utilize many "atomized" services, offered by various institutions and their mutual orchestration, but also to develop specialized dedicated services, which would support the process of spatial data analysis and transformations. In the authors' opinion, services of this type include implementation of statistical algorithms in digital mapping, which allows creating the knowledge base with the use of the "cartographic added value".

2. The role of a map in a website

Usually a user who reaches for geographical, statistical information or who intends to spatially locate events does not expect a form of information transfer other than a map or an atlas. Thus, the cartographic form of transfer meets the user's expectations. This is one of the important prerequisites of the product usefulness. Another positive result of utilizing maps in websites is related to intuitive maintenance and effectiveness of distribution of spatial information.

Due to diversity of data published in geoinformation websites, a map usually plays the role of integrating particular information components. The map is also the basic element of the website composition. Other elements, such as: text fields, queries, functional buttons, illustrations etc., are placed around it. Similarly to other elements of the user interface, a map may be used for operating the website. Its particular components: symbols, labels, specified legend fields, may be connected with specified operations. They can involve starting predefined functions, calling links to other parts of presentation etc. Lastly, a map remains an independent information transfer. Therefore, the competent combination of advantages of a conventional map and a useful application interface, is the basic challenge at the stage of designing the discussed systems.

2.1. Issues concerning website's functionality

The basic assumption concerning functionality (designing the usefulness) of websites is to achieve the required product quality, which is expressed by such properties as: functionality, quality and content timeliness, accessibility and practicality of use. Thus, the widely understood usefulness of a website is the resultant value of above features. It does not result only from the number and development level of functions or options, which characterize a given tool or a thing. As the validating feature, the usefulness is the total value, which specifies the number of benefits the user can get from a product, assuming low inputs of time and labor. In order to achieve the expected usefulness level, clearly specified rules of user centered design (UCD) should be applied considering the user expectations, as well as possibilities and limitations of particular customers (Kowalski, 2006).

Being an element, which constitutes the website interface, a map is the subject of both web and cartographic functionality. The cartographic usefulness covers such features as: the scope of content, the generalization level, cartographic presentation methods, graphical variables, cartographic projection, the legend scope and outlay and auxiliary elements. For the group of interactive maps all functional elements should be additionally considered, which – by analogy to a cartographic network – may be called the map's functional network. Usually, if the basic rules of map editing are not considered, the map readability decreases, its interpretation is more difficult and, as a result, the transfer efficiency is also lower (Kowalski and Mostowska, 2010). The incorrectly prepared functional network may be noticed at every stage of work with a map, leading to unexpected results.

2.2. The functional network of a map in the website

The functional network of an interactive map is wrapped on the entire cartographic image and on marginal elements (according to the conventional terminology). Every active element of presentation within the map area (signature, surface symbol, name, description), in the legend field, in the map frame or in additional toolbars, may cause a specified, internal or external action. Internal actions include, for example, the change of view parameters (zooming, range modifications), the content scope modification (operations on information layers), re-symbolization etc. External operations include hypermedia navigation – evoking hyperlinks to another (text, sound or image) document in the system. Such a high functional potential of a map strengthens its role in information systems, particularly on the web.

The following components may be considered as the **basic elements** of the cartographic user interface (Figure 1):

– a panel of information layers – a map control panel,
– a legend, which is sometimes included in the list of layers in the control panel,
– toolbars, including the bar of basic navigation tools,

Fig. 1. The example of an advanced cartographic interface, applied in the web portal of the Małopolska Spatial Information Infrastructure (geomalopolska.pl)

– a toolbar (or a panel) of selection of cartographic profiles in websites which offer a wide set of data,
– simple text labels.

The following components may be considered as **auxiliary elements** of the cartographic user interface, which directly depend on the scope of website's functionality (Figure 1):

– panels which control functions of the website; usually graphically coherent with the panels of information layers,
– legends of thematic overlays,
– additional toolbars and dialog boxes used in order to configure maps and user's objects (such as navigation map or comments),
– advanced information (text-and-image) labels.

The legend plays a particular role among marginal elements. In the case of interactive visualization it is closer connected with an cartographic image than in the case of conventional maps. It does not only play the role of a pattern of symbols, but it also comments the current application status or presented events (Kowalski, 2003). Additional modules are also necessary, which may be, for example, related to measurements of time and to controlling of animated maps. Special functions of interactive changes of views may be played by a common linear scale: zooming in and out or

changes in information resolution (e.g. size of reference units). Not all elements of a website, which occur together with a map (such as search forms) will be considered as the functional unity. Fields of spatial-and-time navigation and other buttons, which are instrumentally connected with a map, are also map components – particularly when their functionality without a map would be unreasonable.

At present, application map components are those elements, which mostly contribute to popularization of geographical information over the web (Kowalski, 2008). This results from many reasons. Starting from technical opportunities such as increase of web application possibilities, pragmatic features such as independence from specialized applications towards a universal web browser, to such unserious reasons as fashion. Regardless of the reasons, the marriage of cartography and teleinformation benefits both sides.

3. The geostatistical portal development

3.1. Conceptual assumptions and research challenges

Social-and-economic development, which happens at the era of the information society, requires not only the wide access to data and information, but, first of all, development and distribution of tools, which allow transformation of those source data into the useful and applicable knowledge. A good example of such approach is the governmental strategy of national development "Poland 2030", published in the web (http://zds.kprm.gov.pl/node/19), and its more detailed amendment, the Concept of Spatial Management of Poland (http://www.mrr.gov.pl/rozwoj_regionalny/polityka_przestrzenna/ kpzk/aktualnosci/strony/default.aspx). The authors of those documents have formulated the thesis that development of the information infrastructure is not only a technological issue, but, rather, it is the civilization challenge.

In the strategic document the thesis has been formulated that the condition required for fast convergence of the Polish economy is the assumption of the "polarization diffusion" model, which should substitute the "sustainable development" model. This practically means to support "leading and dynamically developing regions", which will become local development "engines". However, the question – **which** agglomerations, municipalities or districts should be considered as those leading centers – remains very interesting. Similarly, it is worth considering which **particular** regions of the country require the diffused inflow of knowledge, technology or investments.

For example, one of development challenges considered in the document "The Concept of Spatial Management of Poland" is the analysis of "the high professional activity and adaptiveness of labor resources". Such analysis requires that data related to the employment and unemployment levels, as well as average salary levels, are considered. Diversification of the Polish labor market, including so-called development clusters (development centers, suburbs, cities, ex-state farms, areas of prevailing low agricultural productivity or rural-and-industrial eras), should be also considered

together with the analysis of participation of particular levels of education and forms of temporal employment. It is obvious that more similar issues may be formulated. However, in order to answer unit questions, advanced spatial data analysis as well as cartographic visualization and **interpretation** of obtained results should be performed – with respect to, for example, the uniformity of spatial distribution of results.

Questions of such type may be answered with an appropriately developed geoinformation structure of spatial knowledge. Following the authors' ideas, the step towards creation of the SKI is the development of a thematic geoportal, which – through the required orchestration of spatial services – will allow analyzing data distributed by various institutions. Utilization of *spatial data mining* techniques for analysis of GUS (The Main Statistical Office) or MRR (The Ministry of Regional Development) data will allow creating the "cartographic added value". This will contribute to development of the knowledge data base to support decision making processes.

At the beginning of this project's implementation it should be remembered that utilization of exploring data analysis techniques in geographic sciences is highly difficult and complex. Many issues which are analyzed using data mining techniques contain multidimensional data, however particular features are usually independent. Spatial data, which describes a four-dimensional space-time, is highly correlated. Analysis of such data requires that the existing relations are considered.

Statistical data is important for development of the information society. At present, such data is distributed by various institutions in various forms and ranges. One of the most important institutions, which distributes statistical data in the web in Poland, is the Main Statistical Office. However, such data is also distributed by other institutions, such as the State Election Commission or the Ministry of Regional Development. Although heterogeneity of data sources and methods of data distribution has some benefits, it may also result in serious difficulties in data access by an average user. The need to create a unified interface, allowing data selection using substantial criteria (such as themes of data, spatial range of data, level of details or temporal coverage) without the necessity to know all data sources, is the first prerequisite to develop an interactive statistical atlas. Its main task would be to become a point of the common access to distributed statistical data, like geoportal.gov.pl plays the similar role with respect to geospatial data.

Apart from data integration through geoinformation services, one of the main tasks performed by the discussed atlas would be cartographic visualization of collected data. Therefore, it should ensure simple visualization tools, allowing the creation of choropleth or diagrammatic maps, as well as selection of colors for those presentations, transparency or the type of reference background. Such presentation, which would follow the rules of cartographic art, would allow better perception of retrieved data by not highly qualified users, thanks to creation of the so-called "cartographic added value". The operation of data visualization itself contributes to better data perception and creates the possibility to draw basic conclusions or to notice regularities, which govern the data, as a result of the correctly selected graphical form of transfer. In order to achieve above results it is necessary to integrate (using for example TERYT codes)

statistical data from various institutions, with geometric data of administrative units' borders, for which they are mostly acquired and aggregated.

The manner of presenting statistical data only in tabular form dominates in existing statistical portals. Not so many provide spatial reference for the data, and even fewer offer partial interactivity in creating thematic visualizations. A short research in this field, conducted by (Fiedukowicz and Gąsiorowski, 2012), confirms the need to develop the concept of geostatistical portal.

At the stage of designing the geostatistical portal, the basic assumption was to utilize – to the maximum possible level – web services organized in the form of an interactive statistical atlas, basing on diversified sources of statistical data and methods of their processing. The idea of the public access to data acquired by various institutions was to be also implemented with the use of an interactive, user friendly cartographic interface of the website (Figure 2). The characteristic feature of such interface would be the richer selection of tools available to the user, comparing to the most popular mapping location websites. The tool set should contain not only elements controlling the cartographic image, but also a wide set of buttons used for controlling tasks of geoprocessing source data (the statistical analyses subsystem).

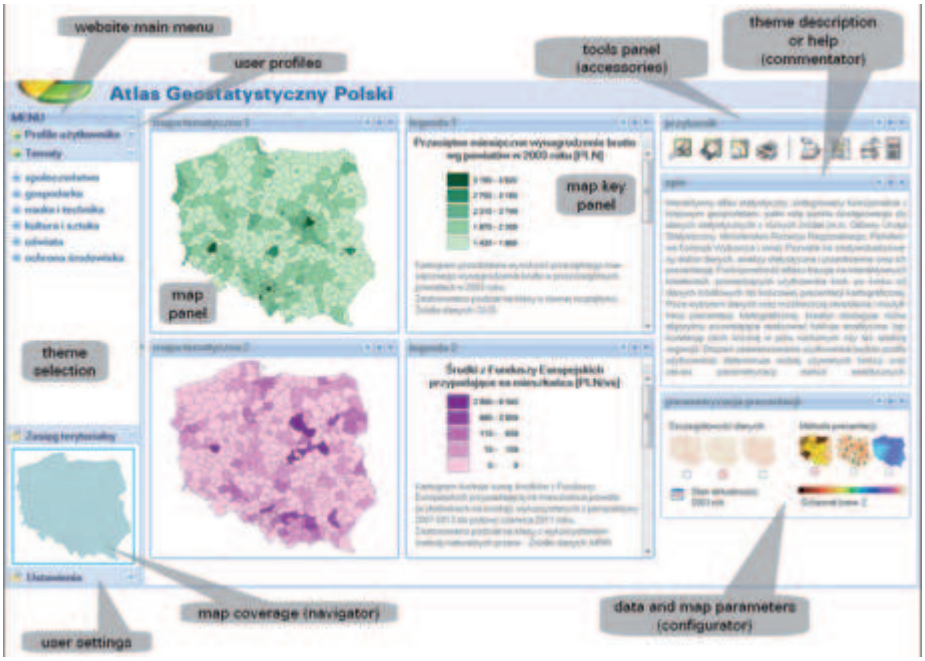

Fig. 2. The design of the User Interface of the Geostatistical Atlas of Poland

An important application component of the designed website would be the creator of thematic maps. It should consider the possibility of independent user selection of such parameters as: data selection, method of geostatistical analyses (e.g. measures of

correlation and regression, movable field etc.) and specification of the cartographic visualization (choropleth map, diagrammatic map, isoline map etc.).

An important aspect of design is consideration of composing schemes, following the rules of cartographic symbolization and generalization. This apparent limitation of the user freedom would guarantee, however, the correct readability and logic of the final cartographic image. At the same time, it would accelerate work for the majority of users who are interested in the final result and not in experimenting with visualization parameters. It will be also possible to utilize pre-made presentations, placed in the basic version of the geostatistical atlas, which do not require the use of the map creator.

3.2. Application components of the developed website

As it was mentioned above, destination of the designed website, different from typical geoinformation portals, will result in modification of some control bars' roles and in development of the user toolbar. In the interface's design there are various fields of lists and fields of selection, which are used for the user interaction with the analyses subsystem. The subsystem of presentation will also answer to the user requests in a modified scope (Figure 3). For example: in a typical map portal the basic operations of the control panel include the change of scale and the level of details of the view (zooming in and out). In the geostatistical website the map scaling mainly influences the change of the information resolution (e.g. the size of data reference units – municipality, district, voivodeship) and not the change of the content.

The information layer panel (the map management panel) – in the case of reference data – is used mainly to switch the visibility of thematic layers (the content change) (Kowalski, 2008). In the designed atlas, the majority of presentations will be limited to several thematic layers. Therefore, tools for modification the ways of presentation and symbolization should be dominating.

The important difference will be noticeable in the scope of the legend field. In many geo-location websites the legend is reduced, displayed optionally or it does not exist at all (Kowalski and Mostowska, 2010). In the geostatistical atlas the presence of the legend results not only from the necessity to explain symbols, but also from the possibility to adapt it to control the presentation. That could involve changing the limits of intervals for steps of measuring scales (choropleth maps, diagrammatic maps). A very important role will be played by the toolbars for selecting cartographic profiles: as a standard – to select a satellite image background or shaded terrain relief. In the case of the discussed project they will be used for the initial organization of the user interface. It will include the basic version of the atlas, the advanced interactive atlas (with the possibility to activate a two- or four-sheet composition for the need of comparing), and the user atlas (with the option of thematic layers creation).

Other tools are listed below:

– tools of names and map descriptions (labeling): simple text labels, advanced information (text-and-image) labels,
– cartometric tools: measurements of distances and areas on a map,

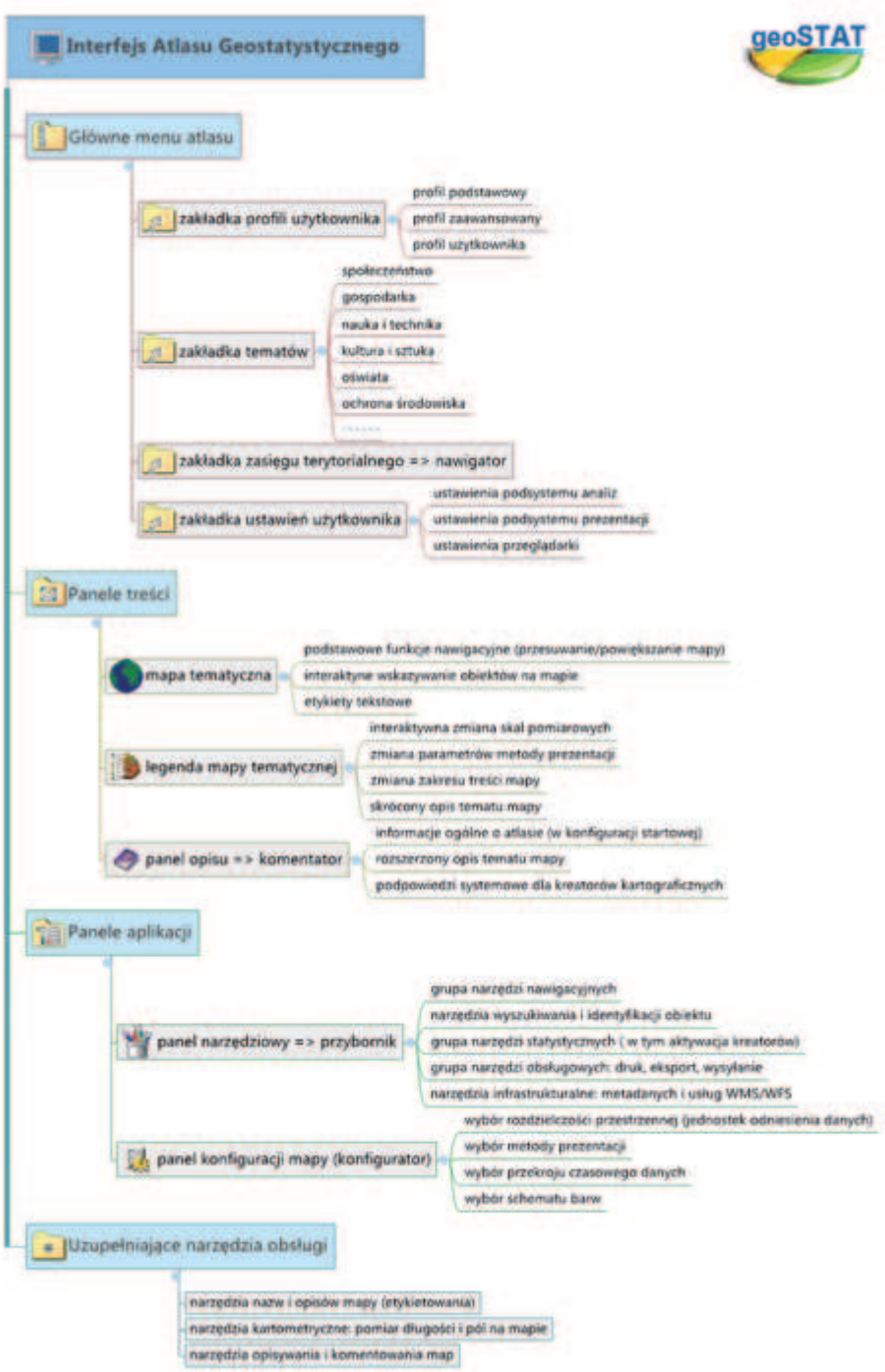

Fig. 3. The functional diagram of the website accessible in the form of the Geostatistical Atlas of Poland

– searching tools according to descriptive or topological criteria, tools of information about the object (object identification tools),
– metadata tools for information files and thematic layers,
– infrastructural tools: adding WMS/WFS services to the user atlas,
– additional toolbars for comments, saving presentation, printing and exporting a map.

3.3. Examples of website's functions and ways of their utilization

Besides outlining the idea of the website and development of its graphical interface, the authors also attempted to practically utilize *data mining* techniques to analyze spatially located statistical data.

The example described below discusses (at the level of a municipality) the percentage of individuals of more than 65 years of age in the population of the Masovian Voivodeship. Three various approaches of presentation of this phenomenon using the statistical geoportal have been proposed. Those approaches are the results of certain statistical analyses, performed for source data, starting from the simplest analysis and ending with the more complex analysis which requires calculations of a series of values and statistical characteristics.

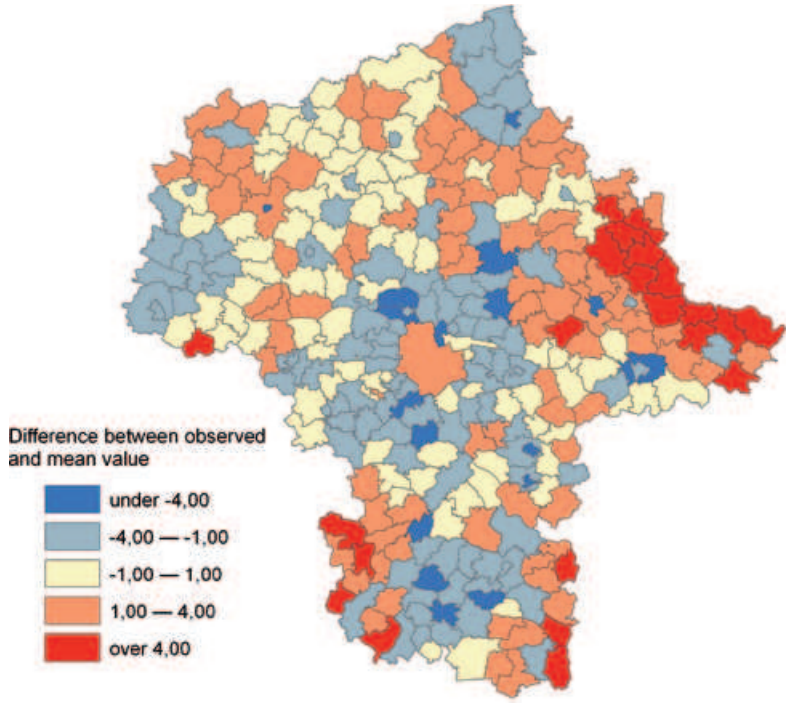

Fig. 4. Percentage of individuals of more than 65 years of age expressed as the difference between values observed in particular municipalities and the mean value

The fist variant assumes presentation of the percentage of people above 65 years of age in the entire population for each municipality, in the form of differences between the observed values and the mean arithmetic value, calculated basing on all observations (for municipalities within the analyzed area). The results of such analysis are presented in Figure 4. This simplest analysis requires that all values of the required attribute are collected by the web service, that the mathematical mean value is calculated basing on values of those attributes, as well as that differences between the mean value and values observed for each municipality are calculated; results are than displayed by the WMS service, with the use of defined symbols.

The second variant is based on the analysis of the influence of a certain factor on the observed percentage of individuals of more than 65 years in the population for each municipality. The discussed factor is the characteristics of a municipality, which classifies it with respect to the urbanization level. Three classes of the municipal status: urban, urban-and-rural, rural, are distinguished. Division of municipalities with respect to their status in the Masovian Voivodeship is presented in Figure 5.

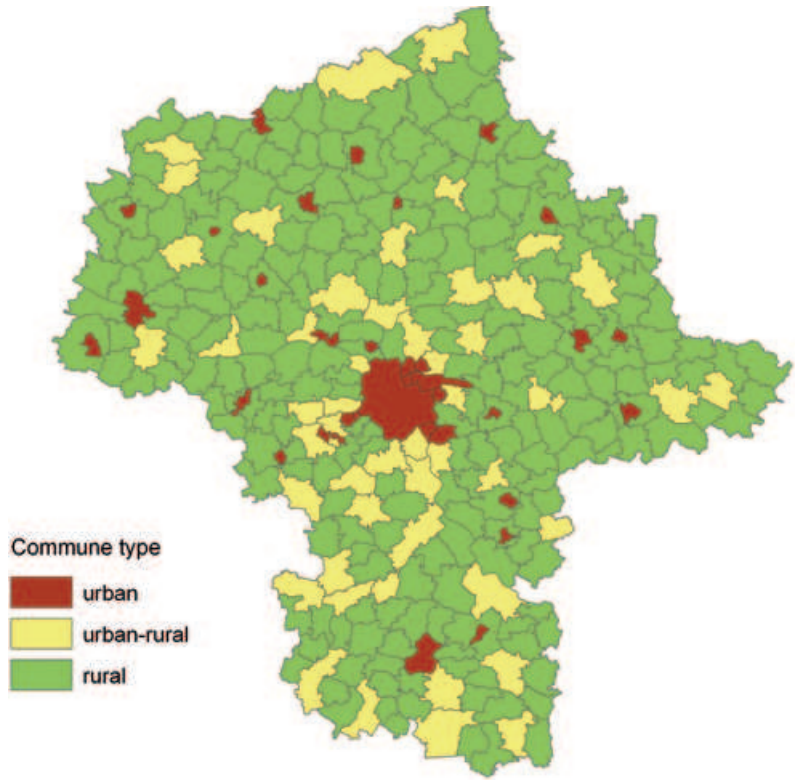

Fig. 5. Municipalities divided by status

This variant requires that all values of corresponding attributes are received by the web service. In this case, it's the attribute representing the percentage of individuals of

more than 65 years of age in the population and the attribute which represents the status of a municipality. Then a series of calculations, required by the single-factor **an**alysis **o**f **va**riance (ANOVA) model are performed (the integer total of squares of deviations, the number of freedom levels, the mean square of deviations, the statistics F). The essence of this analysis is to investigate the importance of differences between mean values calculated on the basis of several groups (three groups in the discussed case), into which observations (municipalities) were divided. The next step should comprise verification of the validity of the hypothesis saying about the influence of the factor (the independent variable – status of a municipality) on the investigated phenomenon (the dependent variable – participation of individuals >65 years). For this purpose the F value should be compared with the limit value, read from the F-Snedecor distribution matrix (existing in the repository) for theassumed level of importance, basing on the levels of freedom of the numerator and the denominator. When the F value is greater than the critical value, calculations of mean values in groups, as well as differences between the mean and observed values will be performed and then the result will be displayed by the web service in the geoportal. In the opposite case, a message should be displayed, informing about the lack of justification for such visualization due to the lack of relations between the variables.

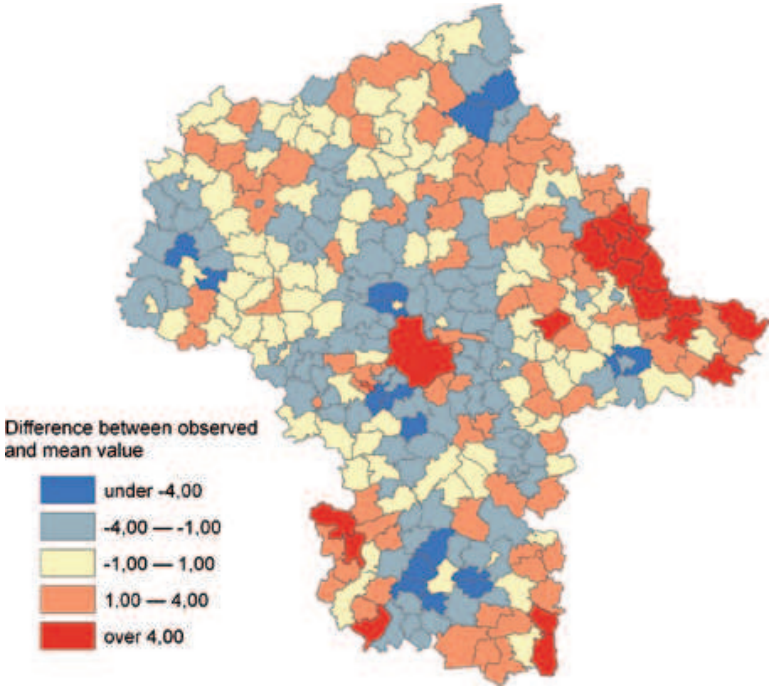

Fig. 6. The diagram of interactions between the percentage of individuals above 65 years of age expressed as differences between the values observed in particular municipalities and mean values for groups determined by the status of a municipality

Visualization of the performed analysis will illustrate the difference between values observed in municipalities and forecasted values (in ANOVA model these will be the mean values in particular groups). Such visualization is presented in Figure 6.

The graphical interpretation of identified interactions between the dependent variable (participation of individuals >65 years), and the independent variable (status of a municipality) is presented in Figure 7.

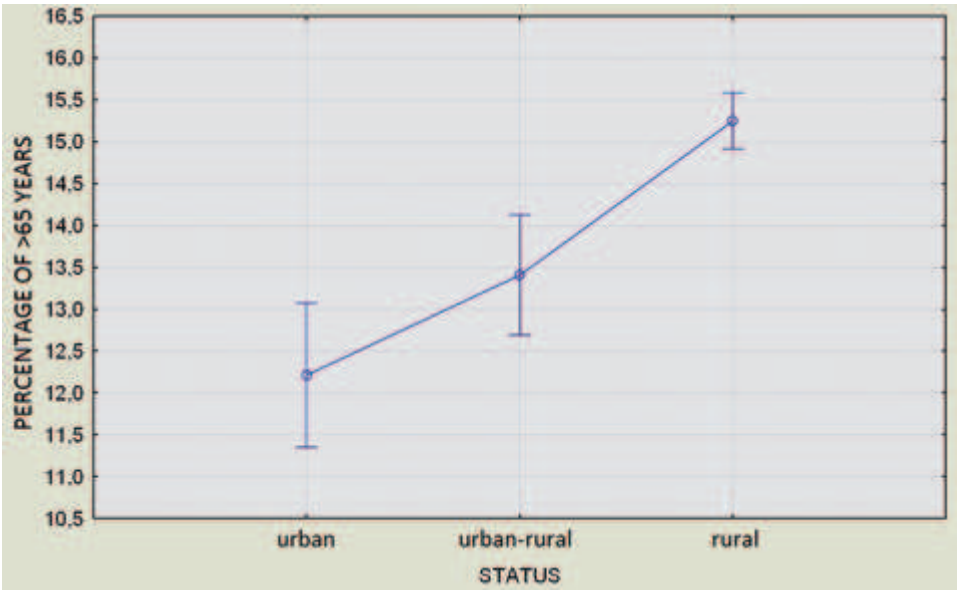

Fig. 7. The diagram of interactions between the percentage of individuals above 65 years of age and the status of a municipality

The third and the most advanced variant of visualization assumes the simultaneous influence of two factors on the dependent variable, which is the percentage of individuals of more than 65 years of age in the population. Similarly to the previous variant, the ANOVA model will be utilized for verification of the hypothesis concerning the simultaneous influence of those two factors; in this case this will be the multi-factor model for permanent results. Besides the status of a municipality, the effect of distances from the administrative center (the Capital City of Warsaw) will be also analyzed. One of three values, which characterize these distances, has been assigned to each municipality in the voivodeship: close, medium, far. Figure 8 illustrates the division of municipalities of the Masovian Voivodeship with respect to distances from the administration center.

The result of this analysis will be visualization presenting differences between values observed in particular municipalities and forecasted values (Figure 9). Similarly to the previous variant, they will be mean values in particular groups. In this case 9 groups will be formed, what results from assumption of all possible combinations (3 groups distinguished for both independent variables).

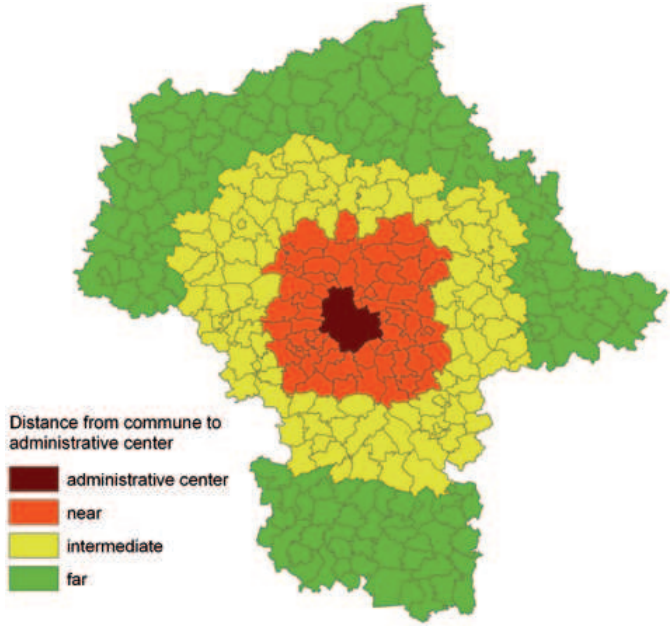

Fig. 8. Municipalities divided by the distance from the administrative center

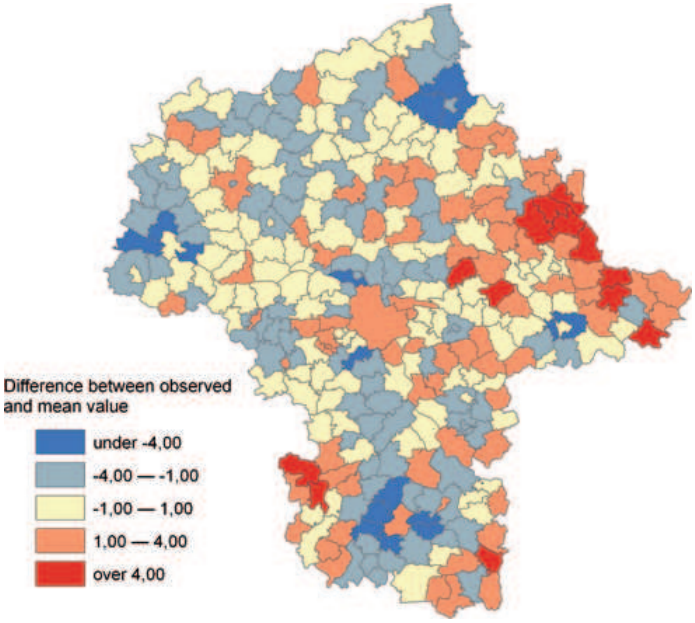

Fig. 9. The diagram of interactions between the percentage of individuals above 65 years of age expressed as differences between the values observed in particular municipalities and mean values for groups determined by the status of a municipality and the distance from the administration center

This variant requires that the web service performs similar operations like in the previous one. However, the level of complexity of those operations will be higher. All values of corresponding attributes must be received. Those include the attribute which represents the percentage of individuals of more than 65 year in the population, the attribute which represents the municipality status and the attribute which represents the distance from the administration centre. The next step will concern calculation of appropriate values for the multi-factor ANOVA model (integer totals of squares of deviations, levels of freedom, mean squares of deviations, F statistics for both independent variables and for interactions). Finally, obtained F values should be compared with critical values and displayed, results should be visualized or the message about the lack of relations between variables should be displayed.

As a result of the performed analysis, differences between mean values, both in groups determined by the status (what has been described above), as well as in groups determined by distances from the administration center and in groups determined simultaneously by both factors, were also noticed. Graphical interpretation of interactions between the dependent variable (participation of individuals >65 years), and the independent variable (distances from the administration center) is presented in Figure 10. The diagram shows that the percentage of individuals of more than 65 years of age in the population increases when the distance from the administration center increases. The graphical interaction between the dependent variable and both independent variables is presented in Figure 11.

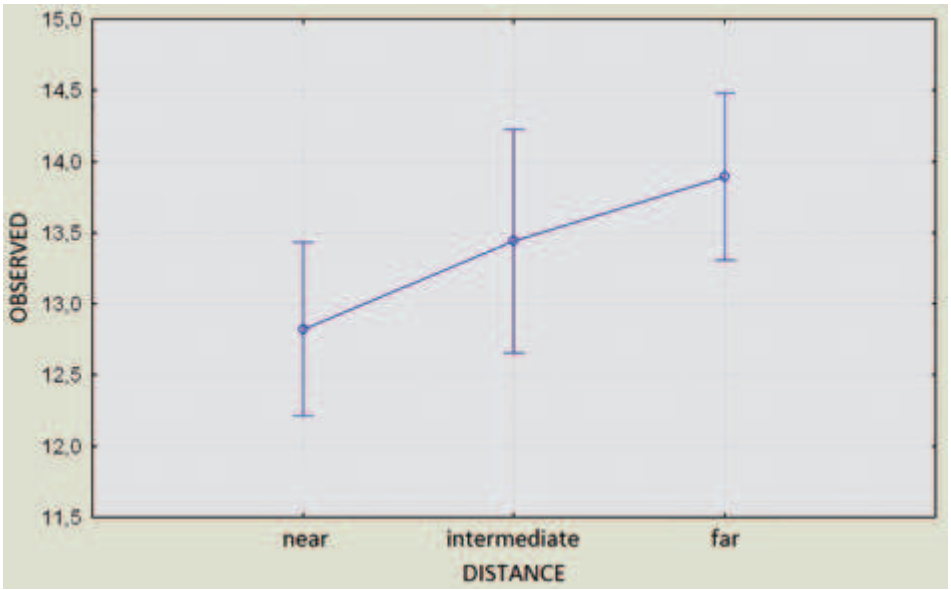

Fig. 10. The diagram of interactions between the percentage of individuals above 65 years of age and the distance between the municipality and the administration center

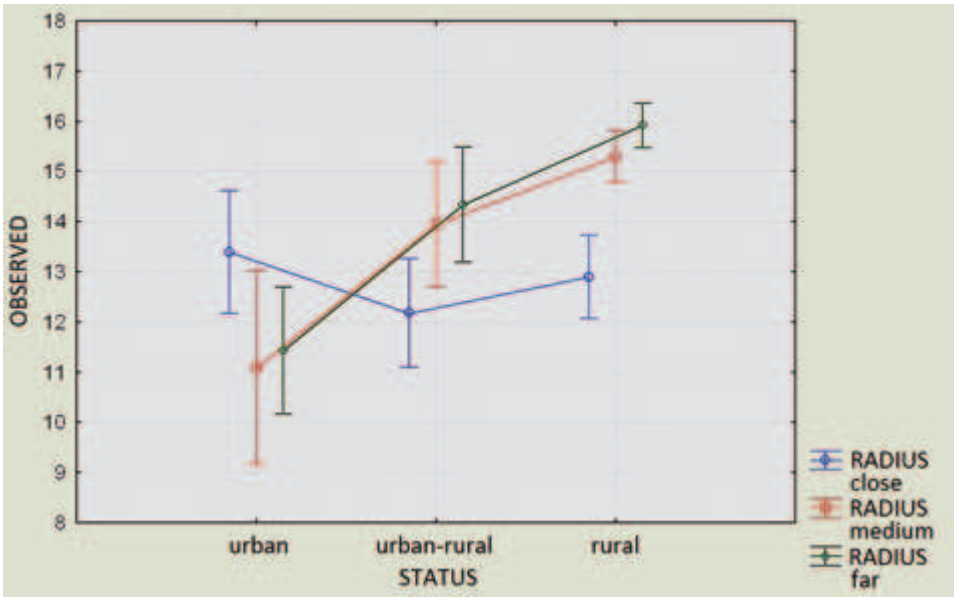

Fig. 11. The diagram of interactions between the percentage of individuals above 65 years of age and the status of a municipality and its distance from the administration center

Certain differences may be observed basing on comparison of visualization of results presented for three proposed variants. In the first case almost raw statistical data was presented, which became the basis for conclusions drawn by the user. In next two variants the same phenomenon was presented, with slight modifications, which resulted in some elimination of the influence of certain trends, identified as a result of analyses performed for the examined area. Random factors of the investigated phenomenon were somehow stressed in those trends; they were stronger stressed in the third variant.

It is worth noticing that in the case of the multi-factor ANOVA model it is possible to test the influence of more than two factors on the dependent variable. Results of analyses of simultaneous influence of three factors (independent variables) on the percentage of highly educated individuals in population of more than 25 years of age, for district, at the country level, are presented below. Those factors are: the district status (distinguished groups: land district, magistrate districts), distance from the district to the closest voivodeship capital (distinguished classes: close, medium, far) and geographic location of the district in the country (distinguished classes: western part, central part, eastern part). These factors are illustrated in Figure 12.

It turned out as a result of the performed variance analysis that all of these three factors influence the percentage of highly educated individuals in the population of more than 25 years of age, at the statistically important level (p = 0.015). Figure 13 presents diagrams, which illustrate relations between the dependent variable and the independent variables.

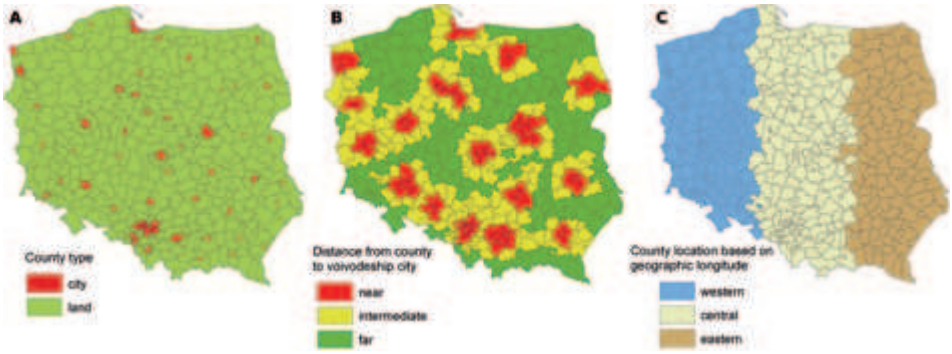

Fig. 12. Factors which statistically influence the percentage of individuals with higher education among individuals above 25 years of age. Division of districts (powiaty) by: A – the status of a municipality; B – the distance from the voivodeship capital; C – geographic location within the country

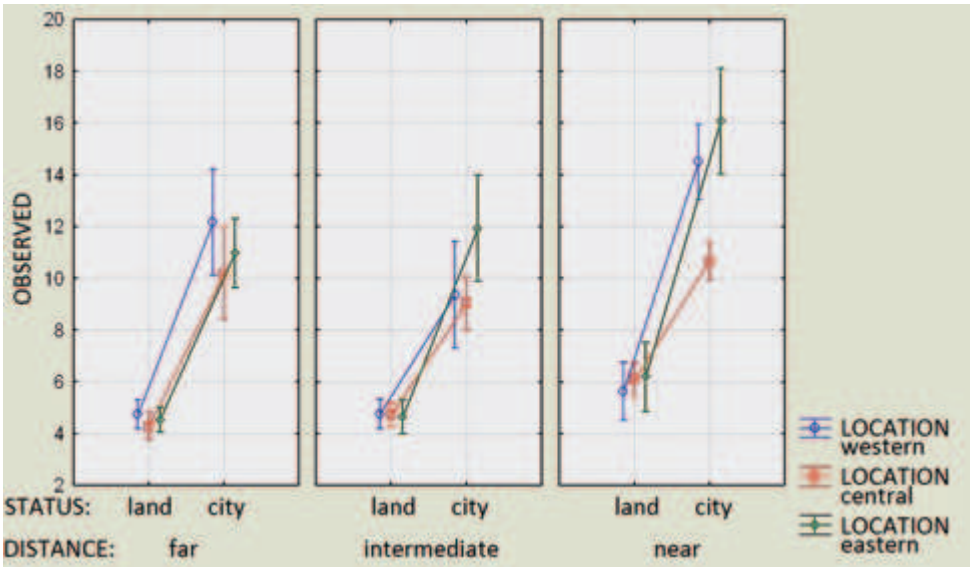

Fig. 13. Diagrams of interactions between the percentage of individuals with higher education among individuals above 25 years of age and the status of a district, its distance from the capital of the voivodeship and the geographic location within the country

It may be stated, basing on these diagrams, that the district status has the strongest influence on the dependent variable. Besides, the differences related to the influence of remaining factors are higher in the city districts than in the land districts. The districts located closer to the voivodeship capitals are also characterized by higher differences caused by the factors of the district status and geographic locations. Due to the fact that important influence of the factors on the dependent variable has been stated basing on comparison of the calculated F statistics with the limit value, resulting from the matrix of the F-Snedecor distribution, 18 mean values, characteristic for particular groups were calculated. Figure 14 presents comparison of two visualization variants of the same phenomenon (the percentage of highly educated individuals in the population of more than 25 years of age). Figure 14A presents the investigated phenomenon in the unprocessed form (similarly to Figure 4). Particularly, clusters of highly educated individuals in bigger cities may be observed here. The Figure 14B, however, illustrates the same phenomenon in the processed way, which eliminates trends and stresses the random factor.

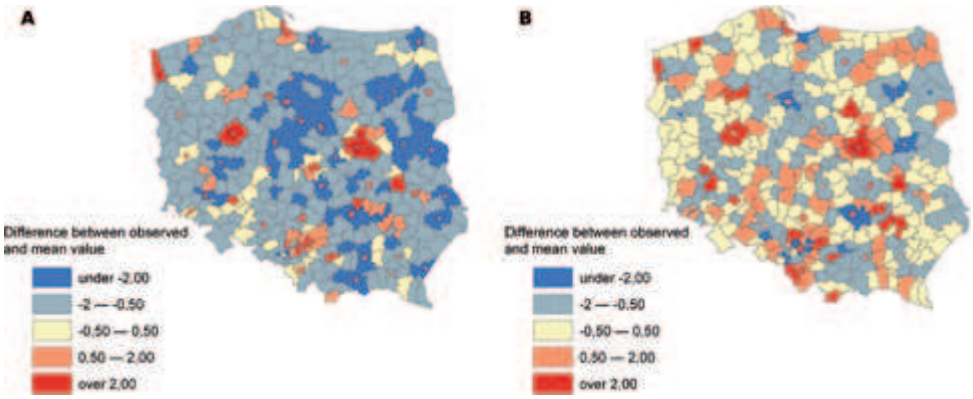

Fig. 14. Diagrams of interactions between the percentage of individuals with higher education among individuals above 25 years of age expressed as differences between the values observed in particular municipalities and: A – the mean value; B – mean values for groups determined by the status of a municipality, the distance from the capital of the voivodeship and the geographic location within the country

For example, the Gorzów Wielkopolski magistrate district in Figure 14A is characterized by the high participation of highly educated individuals in relation to the mean value for the country. However, compared to its own group (magistrate districts, located in western Poland and close to a voivodeship city), it is placed below the average, what may be observed in Figure 14B.

It is obvious that before the analysis of statistical data is started, it should be checked whether requirements concerning implementation of such analysis are met. In the case of the multi-factor ANOVA model, such requirements concern the comparability of variances in groups and the normality of distribution of the dependent

variable in each group. The latter condition may be checked, for example, running the Kolmogarow-Smirnow test. Due to these requirements, the web service should import appropriate data before the given analysis is started, then it should perform calculations and display the message concerning the reasonability of this analysis for specified data.

The above analyses allowed achieving the certain added value, comparing to raw data. However, it would be a mistake to state that results presented in those visualization variants, supported by more advanced processing and analysis are, in general, "better" than the simplest variants. Everything depends on the user and the knowledge, which is to be acquired basing on statistical data presented in the geoportal.

4. Future works

4.1. Directions of planned development of the portal's functionality

The next level of development, which allows creating the added value and which may be implemented with the use of the interactive statistical atlas, is the possibility of statistical analysis of data performed at various levels. Therefore the atlas functionality should allow the use of the synergy effect, related to integration and crossing the data from various sources or data connected with various fields. This effect may be achieved by using simple operations on data (such as division of corresponding values) or by performing statistical analyses of such data pairs. Such analyses include, for example, calculation of correlation or the linear regression model between variables (Figure 15), which allow retrieving the relations mutually combining the data.

Although such operations are statistically correct, they do not consider the spatial aspect of data, besides the element of visualization of result of such analyses. This results from the fact that calculation of residuals from regression (Figure 16) within the administration division units (districts in the discussed example) or correlation values between variables within the movable field (and later interpolation of results) does not consider the neighborhood relations, which occur between those spatial units.

Considering the neighboring districts or movable field as mutually independent, is the unnecessary simplification coming from conventional statistics. Its rejection allows analyzing data in a deeper and more complete way, as well as "drilling knowledge". Weight matrices may be used in the modeling of neighborhood; depending on the assumed model they may take various forms. Regardless of the method, the assumed way of modeling neighborhood may become the better basis for adjustment of the regression model than the conventional statistical methods (Figure 17). Functionality of this type is the proposal for more advanced and aware users. However, in order to ensure better interpretation results of such analyses still should be visualized in the cartographic form.

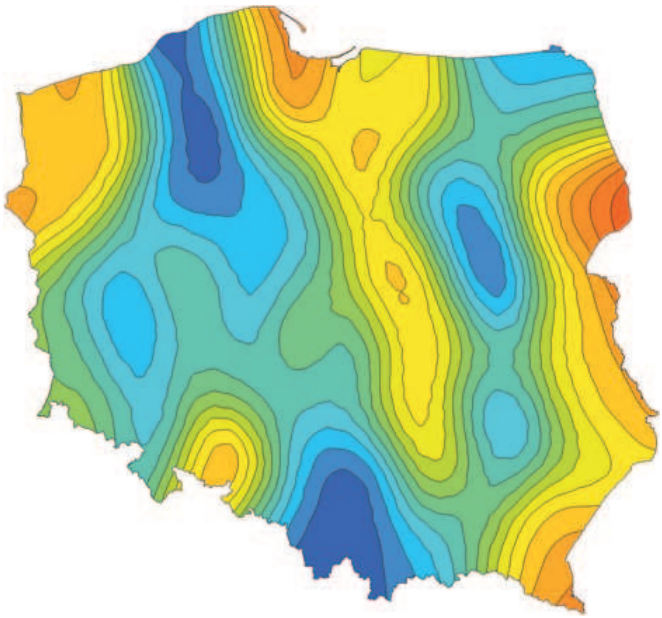

Fig. 15. Isolines of correlation between the percentage support for the Polish accession to the European Union and the distance from the western frontier of Poland (correlation between features, calculated in a movable field)

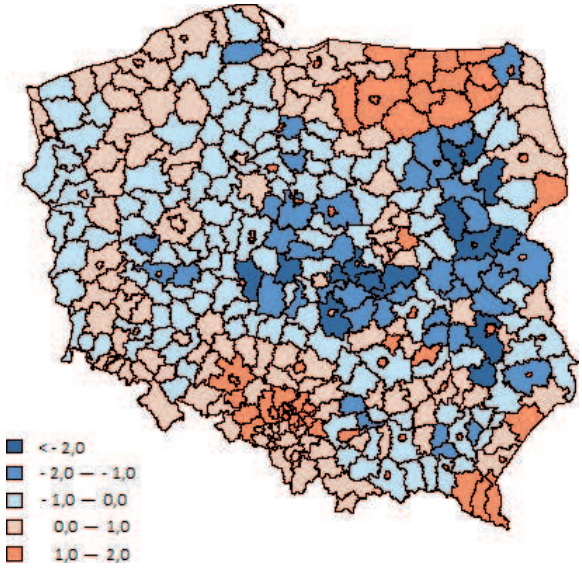

Fig. 16. Standardized residuals from the linear regression model based on the least square methods (the model specifies relations between the percentage support for the Polish accession to the European Union and the distance to the western frontier)

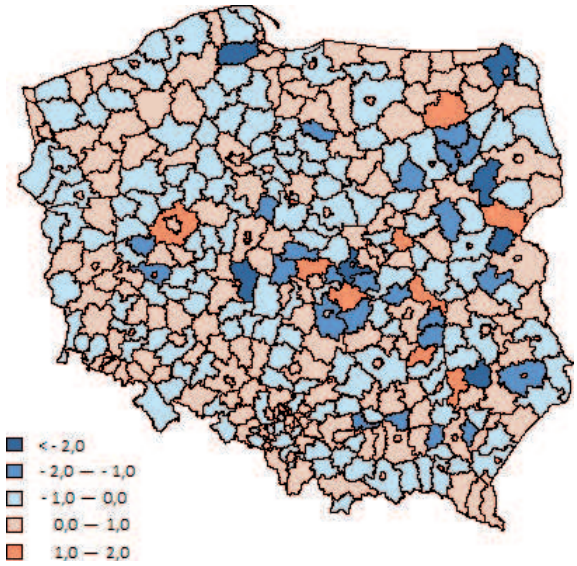

Fig. 17. Standardized residuals from the linear regression model considering the neighborhood of districts based on the model of spatial error, with the use of zero-one matrix of weighs and the presence of the common border of districts (the regression model specifies relations between the percentage support for the Polish accession to the European Union and the distance to the western frontier)

4.2. Testing the usefulness

The new or modified website requires verification and evaluation of its usefulness. Such evaluation is always practically performed by using published data and documents. In the case of the implemented geostatistical portal project, heuristic evaluation will be initially performed by experts, following the guidelines proposed by J. Nielsen (2003). It is inexpensive and easy to perform, and it allows detecting about 80% of design errors on websites. However, testing is the most effective method of checking the functionality (Kowalski, 2006). In the phase of designing, it will be performed by test groups, and after publishing – directly in the user's operating environment. A detailed plan of experiments and the testing station will be prepared. The methodology of testing the functionality is still not popular, although the testing workshop and methodology of research may be adjusted to existing financial possibilities, and results of its utilization eventually lead to measurable benefits.

5. Conclusions

The above functionalities are only the examples of possible services, allowing users to get the added value from information sources which are commonly accessible, although dispersed. Successive levels of development are not only possible, but recommended. Their appearance solely depends on the will and intentions of the developers of the

portal and the analysts, who may propose additional functionalities. For example, the aspect of time in performed analysis (models of diffusion of shocks etc.) could be considered.

All functionalities described above should be performed with an interactive creator. Its complexity must be, however, adjusted to the level of end user capabilities. It should be different (simpler and limited at the same time) for users with the lowest skills, whose operations are limited to importing and visualization of data. Still, it should be more advanced (and supported by detailed methodological description) for researchers, who aim at deeper and more methodological data analysis. Therefore, the proposed interactive statistical atlas should, deliver products and services adapted to every member of the (geo)information society level, independently on his/her level of capability. On the other hand, it should develop the society itself via the educational value, increasing the level of knowledge and involvement of its users.

Acknowledgments

This project is supported by the National Science Centre under the grant No. N N526 043740. The authors would like to thank the anonymous reviewers whose comments materially improved the paper.

References

Berlant, A.M. (1973). Issues of map using theory in scientific research. *Kartograficzna metoda badań w geografii*, IG PAN, 3/4, 39-50.

Berlant, A.M. (1978). *Kartograficzeskij mietod issledowanija (Methods of cartographic research).* Moskwa: Izd. Moskowskogo Uniwiersiteta.

Fiedukowicz, A. Et Gąsiorowski, J. (2012). Cartographical aspects of data mining for knowledge discovery from the agricultural and national census data. (in Polish). *Annales of Geomatics* (In Press).

Kowalski, P.J. (2003). Dynamic cartography – methodological and technical challenges. (In Polish). In Materiały XII Szkoły Kartograficznej (pp. 49-62).

Kowalski, P.J. (2006). Issues of functionality of cartographic presentation in Web information services. (In Polish). In Materiały XXXI Ogólnopolskiej Konferencji Kartograficznej (pp. 103-127).

Kowalski, P.J. (2008). Technical aspects of editing and using geoinformation services. (in Polish). *Polski Przegląd Kartograficzny*, 40(4), 337-348.

Kowalski, P.J. Et Mostowska, M. (2010). Evaluation of the quality and functionality of maps in Web location services. (In Polish). In Materiały XXXIV Ogólnopolskiej Konferencji Kartograficznej: Polska kartografia w dobie przemian metodycznych i technologicznych (pp. 111-127).

Iwaniak, A. (2011). *An intelligent Geoportal*. (in Polish). In III Konferencja z cyklu "Wolne oprogramowanie w geoinformatyce", maj 2011, Wrocław, Polska: Uniwersytet Wrocławski.

Miller, H.J. Et Han, J. (2001). *Geographic data mining and knowledge discovery*. London: Taylor&Francis.

Nielsen, J. (2003). *Designing Web Usability.* Gliwice: Helion.

Olszewski, R. (2006). *Utilization of computational intelligence methods in cartographic modeling.* In at Żyszkowska, W. Et Spallek (Eds). Główne problemy współczesnej kartografii (pp. 49-59). Wrocław: Uniwersytet Wrocławski.

Olszewski, R. (2008). *Has GIS killed cartography?* (In Polish). In Wybrane problemy współczesnej geodezji jako podsystemu informacji przestrzennej. Prace Naukowe, Geodezja, 43, Warszawa: Politechnika Warszawska.

Poole, D. Mackworth, A. Ɛt Goebel, R. (1998). *Computational intelligence. A logical approach*. Oxford University Press.

Saliszczew, K.A. (1955). O kartograficzeskom metodie issledowanija (On cartographic research methods). *Wiestnik Moskowskogo Uniwiersitieta, Sieria fiziko – mat.*, 10.

Wachowicz, M. (2001). *An approach for developing a knowledge construction process based on the integration of GVis and KDD methods*. In: Miller H.J., Han J., Geographic data mining and knowledge discovery. London: Taylor&Francis.

Geoportal statystyczny i „kartograficzna wartość dodana" – tworzenie infrastruktury wiedzy przestrzennej

Anna Fiedukowicz[1], Jędrzej Gąsiorowski[2], Paweł Kowalski[1], Robert Olszewski[1], Agata Pillich-Kolipińska[1]

[1]Politechnika Warszawska, Wydział Geodezji i Kartografii
Zakład Kartografii, Plac Politechniki, 00-661 Warszawa, Polska

[2]Instytut Geodezji i Kartografii,
Zakład Systemów Informacji Przestrzennej i Katastru
ul. Modzelewskiego 27, 02-679 Warszawa, Polska
e-mail: r.olszewski@gik.pw.edu.pl; p.kowalski@gik.pw.edu.pl;
a.pillich@gik.pw.edu.pl; jedrzej.gasiorowski@igik.edu.pl

Streszczenie

Szeroki dostęp do danych źródłowych publikowanych w licznych serwisach internetowych sprawia, iż współcześnie problemem jest nie pozyskanie informacji, lecz umiejętne przekształcenie jej w użyteczną wiedzę. Kartograficzna metoda badań, która od wielu lat służy temu celowi w odniesieniu do danych przestrzennych, zyskuje dziś nowe oblicze – pozwala na wykonywanie złożonych analiz dzięki wykorzystaniu intensywnego rozwoju technologii informatycznych. Znacząca większość zastosowań metod analitycznych tzw. eksploracyjnej analizy danych (*data mining*) i ich „wzbogacania" (*data enrichment*) dotyczy jednakże danych nieprzestrzennych. Wykorzystanie tych metod do analizy danych o charakterze przestrzennym, w tym danych statystycznych, i zapewnienie dostępu do nich w formie dedykowanych usług przyczyniłoby się, zdaniem Autorów, do przetworzenia infrastruktury informacji przestrzennej (*Spatial Information Infrastructure* – SII) w infrastrukturę wiedzy przestrzennej (*Spatial Knowledge Infrastructure* – SKI). Rozwojowi SKI mógłby służyć geoportal statystyczny, którego propozycje funkcjonalności, obejmujące zarówno analizę jak i wizualizację danych, zarysowano w artykule. Zaprezentowano też przykłady analiz statystycznych (ANOVA, regresja z uwzględnieniem sąsiedztwa przestrzennego), możliwych do zaimplementowania w takim portalu, a które mogłyby się przyczynić do wytworzenia „kartograficznej wartości dodanej".