# VISUALIZATION OF STATISTICAL INFERENCES OVER CZSO DATABASE

***Vít PÁSZTO[1], Pavel TUČEK[1,2], Lukáš MAREK[1], Lenka KUPROVÁ[3]***

*[1]Department of Geoinformatics, Faculty of Science, Palacký University in Olomouc, Tř. Svobody 26, 771 46 Olomouc, Czech Republic*

*[2]Department of Mathematical Analysis and Applied Mathematics, Faculty of Science, Palacký University in Olomouc, Tomkova 40, 779 00 Olomouc - Hejčín*

*[3]Department of regional analyses and information services, Czech Statistical Office Pardubice, V Háji 872, 531 53 Pardubice*

vit.paszto@gmail.com, lukas.mar@seznam.cz, pavel.tucek@upol.cz, lenka.kuprova@czso.cz

**Abstract:** There has been more intensive use of GIS in the state statistical services of many national and supranational entities (UN or EU) since the beginning of the new millenium in response to the rapid development of geospatial technologies. This trend has led to the establishment of geographic departments (departments of GIS) within the structure of the principal organs of the state statistical service in many countries. Statistiacal data are from various surveys are interconnected with the space, while their analyses with emphasis on their geographic distribution are the most suitable just for GIS environment. Subsequent visualization of analyses results leads to faster and often more comprehensive insight of the studied issue. This applies to both workers in public administration and the general public.

This paper provides an overview of possible visaulization methods of one and the same dataset and more datasets, which were (geo)statistically analyzed. The paper also shows how GIS tools could be used in the processing of primary data from various registers and surveys of the Czech Statistical Office (CZSO). Furthermore, the concrete example of the CZSO database transformation into proper form useful for GIS processing is also mentioned in the paper. Despite the large number of collected indicators, there are some of them, which are not needed (for GIS analysis) and thus complicate the database structure. On the contrary, some important indicators (e.g. for evaluation of countryside) are not available. Geostatistical tools (whether implemented within GIS or not) are able to derive some of necessary indicators. It is possible to estimate data or indicators via interpolation methods in places where the collection has not been made (or is missing). Following proper visualization of results adds value to overall understanding of investigated issues.

**Keywords:** GIS, Visualization, Statistical Database, Geostatistics, Cartography

## 1. Introduction

The statistical survey has a long and rich tradition in European countries and the Czech countries are no exception. The first general population census in our country took place already in 19th century. Various statistics were not just confined to the population survey over time, but were also targeted other characteristics - gender, marital status, property of people, as well as homes, military sites, etc. Until now, this statistical survey has grown to be available to the general public via many statistical databases and registries (or theirs non-classified parts).

National statistical service in the Czech Republic is carried out by the Czech Statistical Office (hereafter as CZSO). Along with the new concept of Information System of Public Administration (in Czech ISVS) comes the definition of the new four cardinal registers [18] - Register of Territorial Identification and Addresses (in Czech RÚIAN), The Register of Citizens (in Czech ZRO), The Economy register and Estate register (in Czech ZRN).

Nowadays, the CZSO has a number of statistical databases available, whose thematic scope is very wide and thus suitable for geovisualisation. There is a database from the Population and Housing

Census, Register of Economical subjects, Register of Enumeration Districts and Buildings, Register of Accommodation, Database of Foreign Trade, Municipal statistics and more.

As it has already been mentioned, there are geographic divisions or GIS departments of national statistical services around the world, which deal with collecting, managing, analyzing and displaying geospatial data. Subsequent visualization of this datasets is transmitted to the general public. One of the most advanced web applications, which offers graphical presentations of this datasets, is OECD eXplorer (Organization for Economic Co-operation and Development). The application provides tools for analyses of regional statistics of the OECD (and for your uploaded data) using interactive maps, time-series analyses or 2D statistics (Fig. 1).

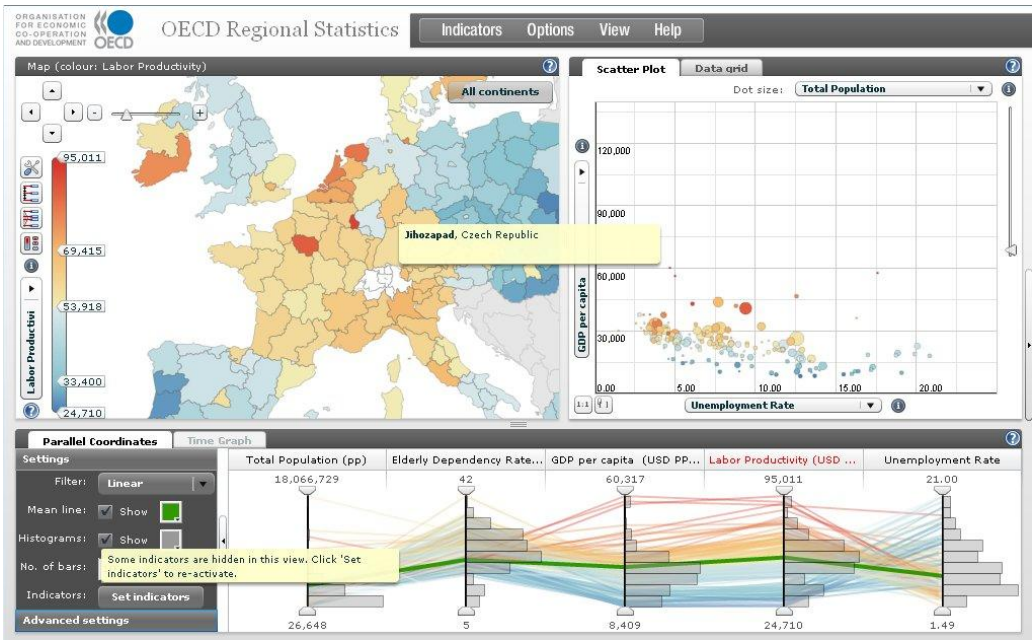


**Fig. 1**: Illustration of OECD eXplorer web-application.

Czech Statistical Office provides a similar application (though not as interactive as OECD eXplorer), which is called Public database (hereafter as PDB), which is described in the following chapter. There will be also mentioned the problem of defining rural space and there will be designed the methodology for its redefinition using fuzzy sets and logic, later in the paper.

## 2. Public database and its geovisualization possibilities

Important source of freely available data is the so-called Public database (PDB). Although the data are already aggregated, they are from various sources (from the CZSO statistical tasks, but also from external sources of statistical service, such as the Ministry of Labour and Social Affairs), covering a wide range of topics − from the environment through population to industry, services and macroeconomics.

However, visualization of data from the PDB via map outputs (you can also choose visualization by graph) is very simple and does not reflect basic cartographic rules (Fig. 2). It is clear that the data visualization using the map form is not the main aim of the Public database. It is just a simple tool for quick information, but often simple displaying of primary datasets is not enough. It is recommended to equip visualization (by map) with the scale bar and give users the chance to select the colour expression of maps (as in OECD eXplorer). It would be also very appropriate to give users the ability to choose the type of classification of the data primary scale-range. Another problem is that there are mostly basic data in the Public database, so that users can make calculations according to their needs. Relative data, which are comparable and strictly used for comparisons by regional statisticians,

are usually not included in the tables of the Public database, so the simple tool for visualization may mislead users to display incomparable data. The Public database is still a developing product in many ways and it cannot be used directly for regional analyses.

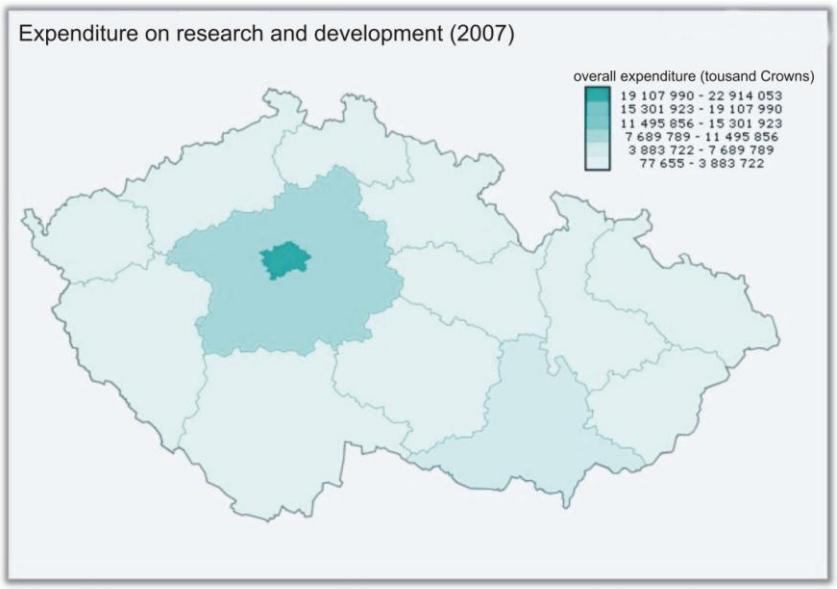


**Fig. 2**: Example of visualization one of indicators from PDB.

Figure 2 shows the distribution of expenditures on the research and development by Regions in the Czech republic, but in absolute figures, not per capita or per regional GDP, which would be more correct. Selected partition of primary scale-range into the equal interval distribution is not appropriate to view the phenomenon properly in the reference map. According the fact that the Czech Republic is divided into 13 Regions and the Capital city Prague, the number of intervals does not follow the rules for displaying geographic phenomena in maps (eg. [4,17]).

Figure 3 shows a sample visualization of the same dataset, but in different colours and, especially, there is individually adjusted width of the intervals to match the cartographic rules. The number of intervals of displayed phenomenon has been also modified and thus there are now four intervals to make the map more legible and not too "empty". The cartographer, geoinformatician and also an expert on the topic or at least a regional statistician must be invited to the process of making relevant and cartographically correct visualization [3,16,17]

If a different method for partitioning the primary scale-range (commonly available in GIS software such as Jenks algorithm) is applied and then is submittted for information loss analysis by calculating the Shannon's entropy [12,14], then the resulting map can be obtained in the form shown on Fig. 3 . In this procedure, the resulting optimal number of intervals (for preservation the ideal amount of information − according to information theory) on the map is five. The original boundaries of the new partition made by the algorithm was not modified.

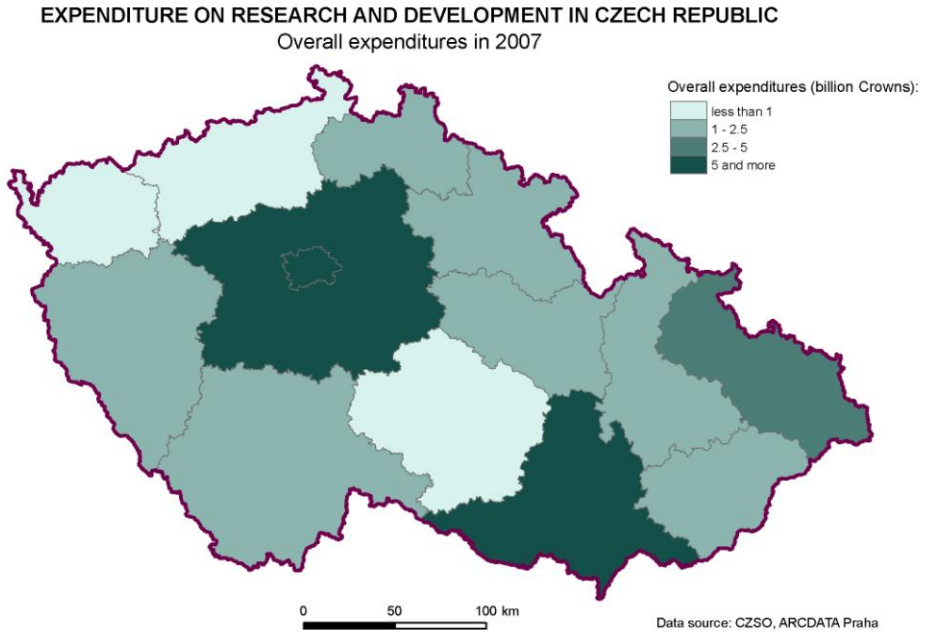


**Fig. 3**: Cartographic visualization of the same indicator from PDB.

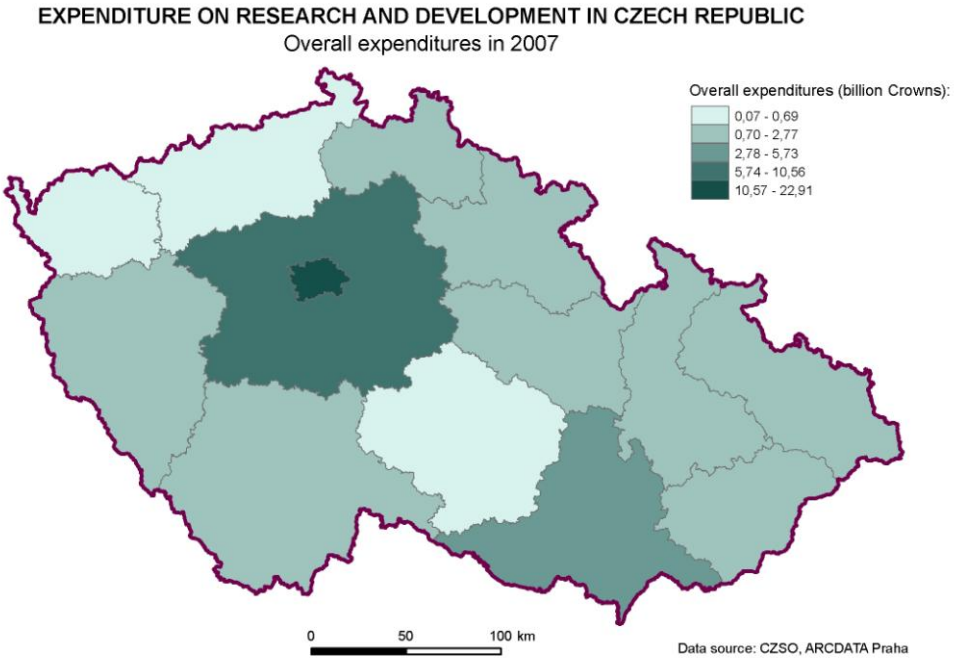


**Fig. 4**: Visualization of the same indicator from PDB according to information entropy analysis.

The most objective method (from approaches mentioned above) of determining the number of intervals of primary data scale-range (which is important for visualization of the phenomenon and its understanding) is the procedure using calculation of the information entropy / information gain. However, it is necessary to note that the final form and design of maps, is involved by a large number of factors (target audience, format and method of an output publication, psychological aspects of users, etc.) and experts from the various topics that contribute to their expert assessment to the final form of visualization of the phenomenon.

It is interesting to note how each of the three images influences the reading of the same dataset and different visualization of phenomenon can be interpreted in several ways. This is one of the elements of a psychological effect of the map on the reader.

## 3. Vizualization of statistical inferences in the definition of rural area

The CZSO's Departments of regional analyses and information services (located in each region) have been solving the problem of defining rural areas using relevant statistical indicators in order to quantify the differences between rural an urban areas on the basis of statistical data. Rural issues have been coming into the force lately.There has been a substantial change in population's movement in recent years due to suburbanization trends. And thus, there is a great distinction between town and country in many aspects of people's lives in each type of area, now in addition with a third type of settlement in semi-urban areas. However, it is not significantly stated how to determine, which village and its surroundings belongs to the urban or the rural area.

There is no uniform definition of rural area these days. The only widely accepted international definition, is the definition from OECD, which is based on the proportion of the population that lives in the territory with a population density of less than 150 persons per $km^2$ [10]. However, this method is now used mainly for international comparisons. The most important document among national papers is Program for rural development in Czech republic (for the period 2007 – 2013). This paper recommends the use of two criteria for the definition of rural area − population and population density. Rural municipality is the village with less than 2,000 inhabitants and a population density of less than 150 persons per $km^2$ [10].

The CZSO's Departments of regional analyses and information services contributed to solve the problem with the definition of rural areas in [11]. They compared 8 variations of the definition of rural areas, ranging from the most simple definitions (such as the statute of the municipality) to a combination of several criteria. The most comprehensive version specified rural areas by multi-criteria evaluation of municipalities and defined another type of classification beyond urban and rural: the transitional type of municipalities (Fig. 5). The specification of the last mentioned type is based on the suburbanization trends in recent years, which was reflected by the criteria used for the definition.

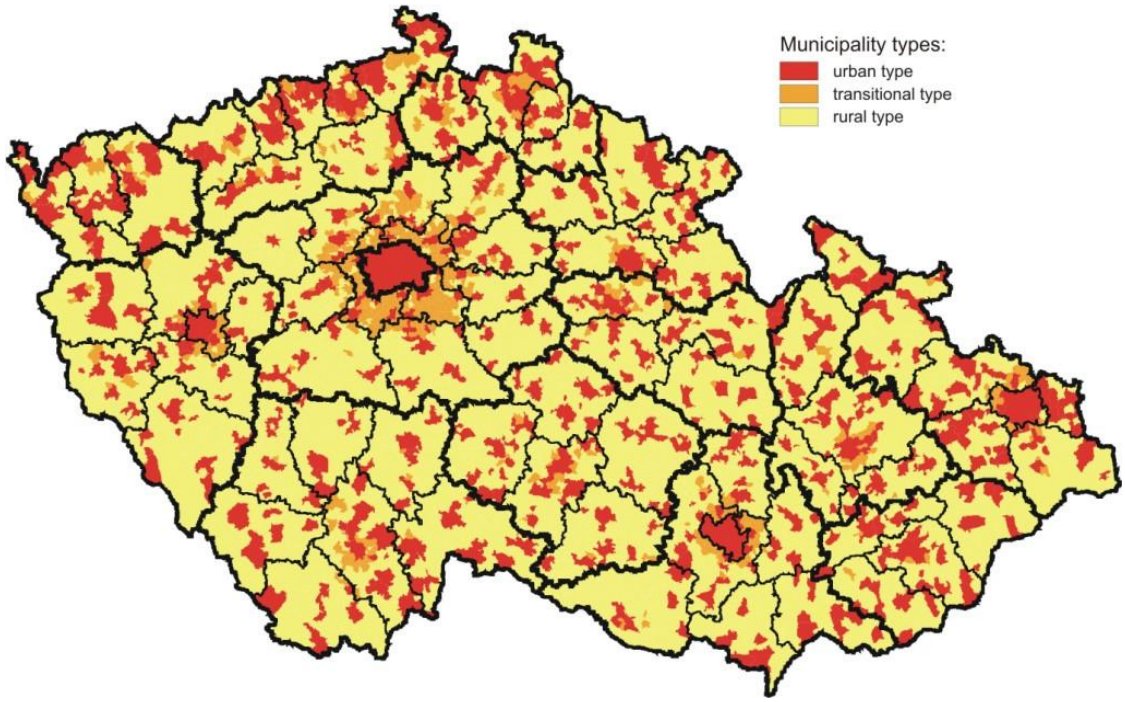


**Fig. 5**: Delimitation of municipality types − red is urban type, orange is transitional type and yellow is rural type. [9]

Proper definition of rural areas is very difficult to compile. There is a lack of significant data (or the data is out-of-date) so there are some aspects statistically hardly formulable. The territorial level of municipalities is too high, their parts are not uniform, and in reality it is often impossible to distinguish between suburban areas of a city and its neighbouring villages. Although the functionality of both types of suburbs (within the city's borders and outside the city) and the life-style of most of their inhabitants is the same. Most definitions of rural areas can operate only with the level of an administrative municipality. The usual definitions of urban and rural areas use fixed and clearly defined indicators. It is then determined that a municipality falls either into the one or the other category. The multi-criteria definition [Venkov] used a system of points assigned to selected ranges of values of the chosen indicators and in addition defined the transitional category, but this did not solve the problem completely.

The following chapter brings a new look into the definition of rural areas. The abrupt transitions between urban and rural type of municipalities can be eliminated using fuzzy logic and fuzzy sets. Although the same indicators are used, operations using fuzzy numbers bring some uncertainty in the definition of rural areas and will smooth the transition between urban and rural areas. With the fuzzy approach and determination of fuzzy weights by experts, municipality can be evaluated such as the 70% belonging to the urban type, 20% belonging to the transitional type and 10% belonging to rural type. This will ensure a smooth transition from the urban municipalities through the transitional type to the rural type (and vice versa). Similarly, it is possible to evaluate each municipality and then visualize it in the map. And by setting different fuzzy weights, the optimum outcome can be achieved for the different aspects of study − demographic, economic, environmental, etc.

## 4. Fuzzy set and logic

Urban and rural areas are naturally defined as thematic structure which is very homogenous. But if we go into more detail we will find that each subgroup of large units is overlapping and even they are thematically related, they provide a large variation of different interpretations, different behaviour or different use. The challenge is therefore how to define the transition, overlay, etc. using the modern theory of fuzzy sets. From a purely mathematical perspective, the transition zone between urban and rural area seems to be a territory where each part can often be, with a certain probability, called as one thematic layer and sometimes with a certain probability as another thematic layer. This approach has led to the formulation of transitional areas between urban area and rural area as a type-2 fuzzy set. As already mentioned above, the concept of a type-2 fuzzy set was firstly introduced in [20] and [19] as an extension of an ordinary fuzzy set (type-1 fuzzy set). Type-2 fuzzy sets have grades of membership that are themselves fuzzy. For more details see [13].

This whole situation can be even more generalized. The application of the theory of fuzzy sets [15] and [5] is introduced as a concept of thematic layers of fuzzy regions. Regions are defined in the theory of geoinformation systems and processes of mathematical modelingas as sharp areas defined by points or polyline. In real situation, however, most regions remain fuzzy. Boundaries of the territory, where characteristics pass from one to another, can not be sharp. A typical example is the object we studied. The transition zone between the urban and rural area, where the transition can not be specified as a single point or line. Fuzzyfication and its use in geoinformation systems and geospatial modeling are discussed with their use in a number of monographs and scientific publications. Example of the use of fuzzyfication in database objects is given in [8].

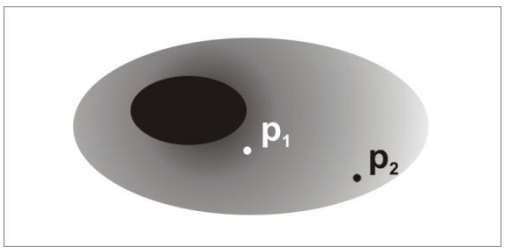


Fig. 6: Graphical illustration of the fuzzy region

Due to the fact that the transition zone is composed of thematic layers, which are often definable as a fuzzy set, the concept of so-called fuzzy regions was introduced [15]. This concept uses the principle of 2D space with the definition of fuzzy sets associated with the membership function in the form:

$$\tilde{A} = \{(p, \mu_{\tilde{A}}(p))\},$$

Where $\mu_{\tilde{A}}: U \rightarrow \langle 0,1 \rangle$ and $p \rightarrow \mu_{\tilde{A}}(p)$.

We have defined the fuzzy set over the domain 2D in this way. Simmilary, if necessary, one can also define a fuzzy set on the 3D field. Operations with fuzzy regions are derived from operations that are defined for fuzzy sets. Their full analysis can be found eg. in [13].

**Modelling of transition area**

Type II fuzzy set is a special type of set, which has not sharply defined the membership function of points to the given set. In the classical case, each element of a defined set has a level of membership function equal to 0 or 1, which indicates whether the element belongs to a given set or not. The degree of membership from the interval <0,1> is assigned to each element in the case of type I fuzzy sets (see [13]). This indicates how the element is identified with a set in which we include it. The generalization of it is even greater in the case of type II fuzzy sets. Transition zone between the urban and rural area is here understood as that set just because it consists of urban and rural areas, where both of the given territories can be defined as a fuzzy set and each element has a certain degree of membership to one of them. More about the application of this type of fuzzy sets can be found in [8].

Quantification of urban and rural areas have always been an aim in many works [2] and [7]. No one has defined the transition between two areas with the use of fuzzy regions. Comprehensive study on the representation of transition zones and defining their boundaries, which can be applied in our case, is given in [5] and [15]. As it has already been mentioned in the introduction, fuzzy theory provides a tool, which can be used for defining the new theorems and definition in the theory of fuzzy regions. If we accept the facts that the representation of urban and rural categories could be done with the use of fuzzy sets and using the fuzzy logic that defines the basic logic operations, we can model the membership function for the transition area as the transition zone. This concept satisfies our requirements.

Let us derive now, how the simple identification of transition areas can be done. Firstly, we will use the α-cut of the fuzzy region. This will define for example the place, which has 0.5 degree of membership to the fuzzy region called “Urban area” or ”Rural area”. The α-cut of the fuzzy region is defined as

$$\tilde{A_\alpha} = \{(p, 1): \mu_{\tilde{A}}(p) > \alpha, p \in U\},$$

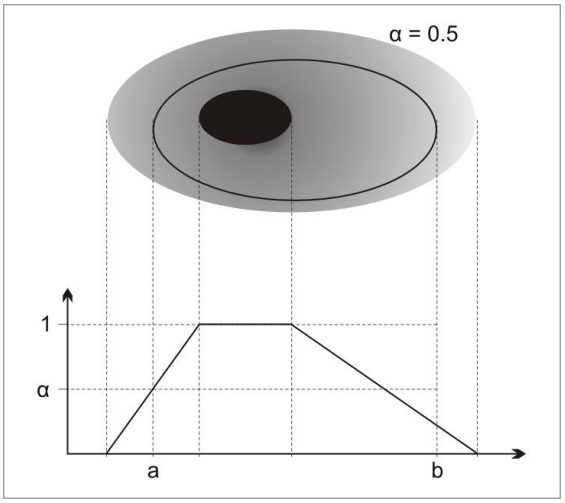


**Fig. 7**: Graphical reprezentation of the α-cut of the fuzzy region.

We will use the intersection operation defined on fuzzy sets. This can be mathematically described as:

$$\tilde{A} \tilde{\cap} \tilde{B} = \{(p, \mu_{\tilde{A}\tilde{\cap}\tilde{B}}(p)) : \mu_{\tilde{A}\tilde{\cap}\tilde{B}}(p) = T(\mu_{\tilde{A}}(p), \mu_{\tilde{B}}(p))\},$$

Where T is a T-norm [6]. The result of the intersection could be seen on figure 8. One can clearly see from the figure that the transitional zone between urban and rural areas are such points that belong to one or second-forming zone with less than 0.5 value of membership function. It follows that the transition zone contains a relatively low membership levels. This can be enhanced by the normalization in order to avoid the misinterpretation. This was the idea, where we have clearly defined the transitional zone between the two functional areas. Authors are now working on implementation of this method to the geospatial analysis of the above mentioned dataset. The result will be visualized as set of fuzzy regions, which will be called „Urban area“, „Rural area“ and „Transition zone “.

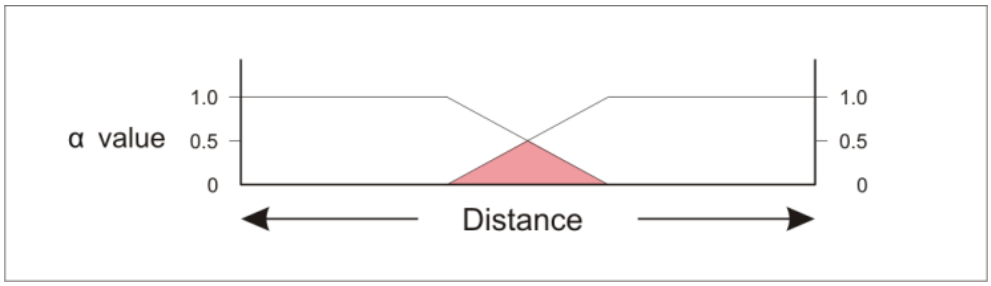


**Fig. 8**: Graphical reprezentation of the α-cut of the fuzzy region.

## 5. Summary and future outlook

This paper provided an overview of possible visaulization methods of one and the same dataset and more datasets, which were (geo)statistically analyzed. The paper also showed how GIS tools could be used in the processing of primary data from various registers and surveys of the Czech Statistical Office (CZSO). Furthermore, the concrete example of the CZSO Public database transformation into proper form useful for GIS processing is also mentioned in the paper. A brand new theory of modelling transitional areas is shown at the end of the article. This theory is based on the application of fuzzy sets, fuzzy regions and fuzzy logic. Due to the fact that the fuzzy theory offers a wide range of application of such sets, we can also clearly define the concepts of the smallest sharp set including the fuzzy region (which can be understood as the influence of urban, or rural parts), convex cover (the city as a whole part with all links) or a range of other concepts based on the algebraic properties of the fuzzy objects.

*Acknowledgement: This work was supported by the grant of Czech Science Foundation no. 205/09/1159 – The Intelligent System for Interactive Support of Thematic Map Design.*

## References

[1] ARNOT, Charles, FISHER, Peter. *Mapping the Ecotone with Fuzzy Sets*. In MORRIS, Ashley, KOKHAN, Svitlana. (eds.) Geographic Uncertainty in Environmental Security. [s.l.] : [s.n.], 2007. s. 19-32. ISBN 978-1-4020-6436-4. ISSN 1871-4668.

[2] FREY, W.H., ZIMMER, Z. (2001): *Defining the City*. In: Paddison, R. ed.: Handbook of Urban Studies. Sage, London, str. 14-36.

[3] KAŇOK, J. *Tematická kartografie.* Ostrava : Ostravská univerzita, 1999. 318 s.

[4] KAŇOK, J. *Kvantitativní metody v geografii (Grafické a kartografické metody).* Ostrava : Ethics, 1992. 233 s.

[5] Kilianová, H., Pechanec, V., Lacina, J., Halas, P.: *Ekotony v současné krajině*, Vydavatelství UP Olomouc, In print.

[6] Klir, G.J., Yuan, B.: *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, New Jersey 1996.

[7] MAYER, H.M. (1971): *Definitions of „City"*. In: Bourne, L.S. (1971): Internal Structure of the City: readings on space and environment. Oxford University Press, Toronto, str. 28-31.

[8] MORRIS, Ashley, KOKHAN, Svitlana. (eds.) *Geographic Uncertainty in Environmental Security.* [s.l.] : [s.n.], 2007. s. 19-32. ISBN 978-1-4020-6436-4. ISSN 1871-4668.

[9] ŠTĚPNIČKA, J., KUPROVÁ, L. *Vymezení venkova na základě multikriteriálního hodnocení*. [s.l.] : [s.n.], 2008 tisk. 2 s. Manuskript.

[10] Český statistický úřad. *Postavení venkova v Pardubickém kraji*. [s.l.] : [s.n.], 2009. 157 s.

[11] Český statistický úřad. *Varianty vymezení VENKOVA a jejich zobrazení ve statistických ukazatelích v letech 2000 až 2006*. [s.l.] : [s.n.], 2008. 23 s.

[12] SHANNON, C.E. A *Mathematical Theory of Communication*. Bell System Technical Journal. 1948, no. 27, s. 379-423, 623-656.

[13] Talašová, J.: *Fuzzy metody vícekriteriálního hodnocení a rozhodování*. Vydavatelství UP, Olomouc, 2003, 180 s., ISBN 80-244-0614-4.

[14] TUČEK, P., PÁSZTO, V., VOŽENÍLEK, V.: *Regular Use of Entropy for Studying Dissimilar Geographical Phenomena*, Geografie, 2/2009, pp 117-130, 2009.

[15] VERSTAETE, Jörg, HALLEZ, Axel, DE TRÉ, Guy. *Fuzzy regions: Theory and Applications*. In MORRIS, Ashley, KOKHAN (eds.), Svitlana. Geographic Uncertainty in Environmental Security. [s.l.] : [s.n.], 2007. s. 1-17. ISBN 978-1-4020-6436-4. ISSN 1871-4668.

[16] VOŽENÍLEK, V. *Cartography for GIS – geovisualisation and map communication*. Univerzita Palackého v Olomouci, Olomouc, 2005, 140 s.

[17] VOŽENÍLEK, V. *Aplikovaná kartografie I – tematické mapy*. Univerzita Palackého v Olomouci, Olomouc, 2001, 187 s.

[18] VOŽENÍLEK, V. *Geoinformační aspekty státní informační politiky ČR*. 1. vyd. [s.l.] : [s.n.], 2009. 187 s.

[19] Zadeh, L. A.: *Fuzzy sets,* Information and Control, vol. 8, pp. 338-353, 1965.

[20] Zadeh, L. A.: *The concept of a linguistic variable and its application to approximate reasoning - 1,* Information Sciences, vol. 8, pp. 199-249,1975.