# OUTLIERS DETECTION IN THE URBAN AREAS. CASE STUDIES: ROMANIA, BULGARIA AND HUNGARY

**Alexandra Sandu**[1]

**Abstract.** Outliers are values that are questionable because they are much bigger or smaller than the majority of the values analyzed. The outlier detection is a research domain about which not a lot of persons are aware of, and because of that the identification of the abnormal values is often a skipped step in the analysis of a database. However, these atypical values which are generally ignored could result into interesting researches which could reveal valuable information about a particular area or about a particular phenomenon studied, exposing patterns which otherwise would have not been observed. Yet, a solid research should be done before removing the observations which were identified as being atypical and this paper emphasizes the duality of the term outlier and the necessity of performing an identification based on more than one method, before deciding if a suspicious value should or should not be eliminated from a research.

***Keywords:*** *outlier, index of Moran I, atypical value, GMES Urban Atlas*

## 1. Introduction

The outliers detection is often a neglected step in the analysis of the databases, their quality control being done in the most of the cases visually, superficially and because of this fact it is quite difficult to identify possible errors that could alter the final results of a research study.

A rapid scan of some research papers reveals that in the most of cases, the authors do not go through steps which could allow them to identify potential outliers, which indicates the fact that the outliers detection problem is still a research domain of which not a lot of persons are aware of and thereby is not valorized as it should be.

The literature on the outliers detection reveals us a variability for the term of outlier as it follows:"An outlier is considerate to be a data point that is far outside the norm for a variable" (Jarrell, 1994; Rasmussen, 1988; Stevens, 1984). Hawkins defines an outlier as being an observation that "deviates so much from other observations that may arouse suspicion that it was generated by a different mechanism" (Hawkins, 1980). Also an outlier can be seen as an observation which is "dubious in the eyes of the researcher". (Dixon, 1950)

However, once the outliers were identified, the problem which occurs is if these atypical observations should automatically be considered as being wrong values and also if we

[1] "Al.I.Cuza" University of Iasi, Faculty of Geography and Geology, Department of Geography, Bd.Carol I 20A, 700505, Iasi, Romania, alexandra_sandu_fr@yahoo.fr

should take into consideration their exclusion from the research study. A possible answer it is found by analyzing the francophone literature on the outliers detection which use in order to define the term of outlier the concept of „valeur extrême", as well as the concept of „valeur aberrant", capturing the duality of the notion of outlier, as it might be an aberrant value, so a wrong value, as well as an extreme value that is not necessarily a wrong value but it is atypical due to exceptional localization conditions.

## 2. Database and methodology

Martin Charlton, whose research focused on proposing new methods for the detection of an outlier, one of his methods being used even in these case studies, proposes a definition that summarizes the majority of the definitions mentioned above. Therefore he defines an outlier as: "an atypical observation, an observation that does not fit the general pattern of the others values which were analyzed".

The study zone was represented by Romania, Bulgaria and Hungary, more precisely the research analyzed 14 urban areas from Romania, 8 urban areas from Bulgaria and 9 urban areas from Hungary.

The database required for this study was taken from the GMES Urban Atlas developed by the European Environment Agency which contains maps that analyze land use and land cover for Large Urban Zones with more than 100.000 inhabitants as defined by the Urban Audit. This study worked with the following categories:

*a) Continuous Urban Fabric (CUF)* - built-up areas with an average degree of soil sealing greater than 80%; predominant residential use, independent of their housing scheme. (Single-family houses family houses or high rise dwellings, city center or suburb).

*b) Green Urban Areas (GUA)* - public green areas with a predominant recreational use (e.g. gardens, zoos, parks)

*c) Industrial, commercial, public, military and private units (ICPMPU)* – the most of the surface is covered by artificial structures (e.g. buildings) or artificial surfaces (e.g. concrete, asphalt or otherwise stabilized surface, e.g. compacted soil, devoid of vegetation) and the land is used for industrial, commercial, public, military or private activities.

This study detected the outliers only for the areas of the three categories mentioned above because the research was conducted on the idea that in the perspective of the current trend of the economic and social development, these categories represent the classes which attract the majority of the investments and therefore they are the main territories that tend to extend from a city, thereby representing the main topics for the research studies focused on the urban areas.

In order to detect the outliers the software ARCGIS 9.3 was used, more exactly the method which is based on the index of Moran (Anselin Local Moran's I), a statistical index for spatial autocorrelation, which measures the similarity between a variable and its neighbors using spatial weight matrices. The outliers may thereby be identified using the following five different methods:

a) *Inverse Distance*: the impact of one feature on another feature decreases with distance;

b) *Inverse Distance Squared*: same as Inverse Distance, but the impact decreases more sharply over distance;

c) *Fixed Distance Band*: everything within a specified critical distance is included in the analysis and everything outside the critical distance is excluded;

d) *Zone of Indifference*: a combination of Inverse Distance and Fixed Distance Band. Anything up to a critical distance has an impact on the analysis. However, once that critical distance is exceeded, the level of impact quickly drops off;

e) *Polygon Contiguity*: the neighbors of each feature are only those with which the feature shares a boundary. All other features have no influence;

Of all the five methods mentioned above, in this paper only the first 4 methods were used because the fifth is based on the detection of outliers for the areas which got a boundary and in this case, all the three categories analyzed (residential, industrial and green urban areas) are separated by the road network.

The index of Moran I is defined by statistics as being a measure of the degree of the spatial autocorrelation of neighbors areas. The formula proposed by Patrick A. Moran does not use only one spatial dimension, but it is multidimensional and multidirectional, which allows the realization of a complex analysis.

$$I = \frac{N}{\sum i \sum j\, \omega ij} \frac{\sum i \sum j\, \omega_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum i\, (X_i - \bar{x})^2}$$

where: N - the number of spatial units indexed by i and j;

X - the variable of interest;

$\bar{X}, \bar{x}$ - the mean of X;

$\omega_{ij}$ - an element of a matrix of spatial weights;

The atypical values which results after applying the four methods mentioned above might have a negative spatial autocorrelation (the neighbors have opposite values) or it might have a positive autocorrelation (neighbors have similar values), while a value close to 0 indicates in the most of the cases the absence of a spatial autocorrelation . Thus, it results four different cases which were individualized according to the difference of the values existing between the variables of the analyzed area represented in Figure 1 by the axis O(x) and the variables of the neighboring areas represented in Figure 1 by axis O(y) as it follows:

a) *Low – High* **(LH)** – a variable which has a low value surrounded by variables which have high values

b) *High – Low* **(HL)** – a variable which has a high value surrounded by variables which have low values

c) *Low – Low* **(LL)** – a variable which has a low value surrounded by variables which also have low values

d) *High – High* **(HH)** – a variable which has a high value surrounded by variables which also have high values

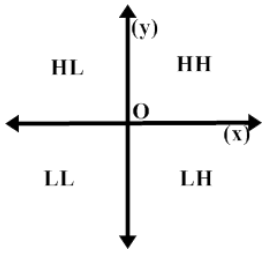

*Figure 1: Types of outliers*

For example, if the area of a residential zone (CUF) has a small absolute value and the neighboring zones have the absolute values of the areas higher than this atypical value it will be classified as a Low-High outlier.

### 3. Results

Using four different methods revealed a versatility of the theoretical dimension of the term of outlier , as well as a spatial variability of the outliers depending on the method used to identify them.

Therefore, for urban areas in Romania, the method which identified the biggest number of outliers is Zone of Indifference, but it had to be taken into account the fact that the method Fixed Distance Band identified in the most cases the same number of outliers as Zone of Indifference, rarely being some insignificative differences. In contrast, the smallest number of outliers was identified by the Inverse Distance Squared method.

The same pattern of outliers detection is found in the cities from Hungary, as well as for those from Bulgaria, the opposition between the biggest number and the smallest number of outliers detected being maintained between the methods Zone of Indifference & Fixed Distance, and the Inverse Distance Squared method. (see Table 1)

Table 1 : Bulgaria – The number of outliers identified depending on the category analyzed and the method which was used for the analysis

| Country | City | The analyzed category | Inverse Distance | Inverse Distance Squared | Fixed distance band | Zone of indifference | Total Outliers | The analyzed category | Inverse Distance | Inverse Distance Squared | Fixed distance band | Zone of indifference | Total Outliers | The analyzed category | Inverse Distance | Inverse Distance Squared | Fixed distance band | Zone of indifference | Total Outliers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bulgaria | *Sofia* | **CUF** | 1292 | 345 | 1983 | 1986 | 5606 | **GUA** | 18 | 17 | 18 | 18 | 71 | **ICP** | 120 | 67 | 238 | 238 | 663 |
| Bulgaria | *Plovdiv* | **CUF** | 910 | 157 | 1405 | 1408 | 3880 | **GUA** | 3 | 3 | 6 | 6 | 18 | **ICP** | 11 | 8 | 15 | 15 | 49 |
| Bulgaria | *Varna* | **CUF** | 946 | 160 | 1412 | 1413 | 3931 | **GUA** | 1 | 2 | 1 | 1 | 5 | **ICP** | 38 | 28 | 64 | 64 | 194 |
| Bulgaria | *Burgas* | **CUF** | 249 | 83 | 279 | 279 | 890 | **GUA** | 4 | 0 | 1 | 1 | 6 | **ICP** | 21 | 15 | 32 | 32 | 100 |
| Bulgaria | *Pleven* | **CUF** | 168 | 29 | 414 | 414 | 1025 | **GUA** | 5 | 2 | 7 | 7 | 21 | **ICP** | 16 | 10 | 21 | 21 | 68 |
| Bulgaria | *Ruse* | **CUF** | 246 | 65 | 427 | 427 | 1165 | **GUA** | 2 | 3 | 2 | 2 | 9 | **ICP** | 42 | 26 | 50 | 50 | 168 |
| Bulgaria | *Vidin* | **CUF** | 96 | 41 | 167 | 167 | 471 | **GUA** | 2 | 3 | 3 | 3 | 11 | **ICP** | 23 | 14 | 26 | 26 | 89 |
| Bulgaria | *Stara Zagora* | **CUF** | 72 | 35 | 38 | 38 | 183 | **GUA** | 2 | 2 | 3 | 3 | 10 | **ICP** | 4 | 3 | 3 | 4 | 14 |

In terms of a classification using the relative values (percentage) of the total number of outliers detected by the application of the four methods, depending on the type of the analyzed zone, we note that for both the residential areas (see Fig. 2) and the commercial and industrial areas (see Fig. 3), the most atypical values were identified by the methods Zone of Indifference and Fixed Distance Band.

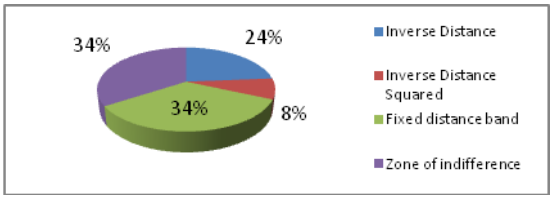


*Figure 2: Continuous Urban Factor – the percentage of the outliers depending on the method used for detection*

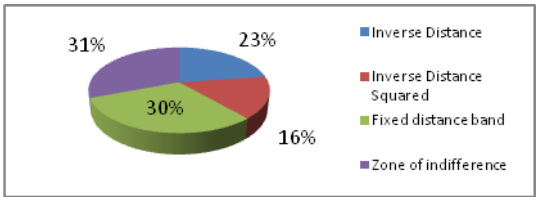


*Figure 3 : Industrial, commercial, public military and private units - the percentage of the outliers depending on the method used for detection*

The urban green areas show some differences from the pattern of the other two areas mentioned above because it may be observed a relative homogeneity in terms of the share of each method used for identifying the outliers. But if it is taken into account that the number of green spaces is considerably smaller than the number of residential, industrial and commercial areas from every city analyzed, it can be said without a doubt that the previously established pattern in terms of the ranking of the methods depending on the number of outliers identified it is maintained in this case, too. (see Fig. 4)

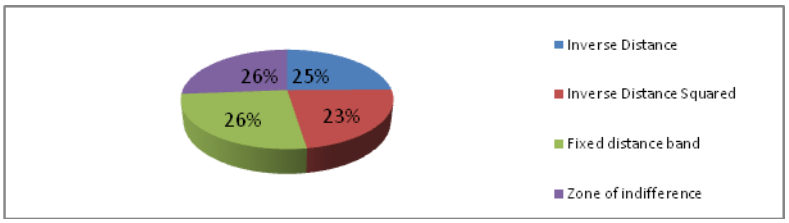


*Figure 4: Green Urban Areas – the percentage of the outliers depending on the method used for detection*

Also a detailed analysis of two maps that shows the differences in the detection of the outliers for the city of Bucharest using the method Zone of Indifference (see Fig. 5), as well as the method Inverse Distance Squared (see Fig. 3), reveals the spatial variability of the outliers and in the same time it emphasizes the need on the one hand to perform a quality control of the databases which will be used in a research study and on the other hand it shows the large conceptual dimension of the term outlier, which can give us an idea of the difficulty in deciding when and on which premises we can identify these atypical values.

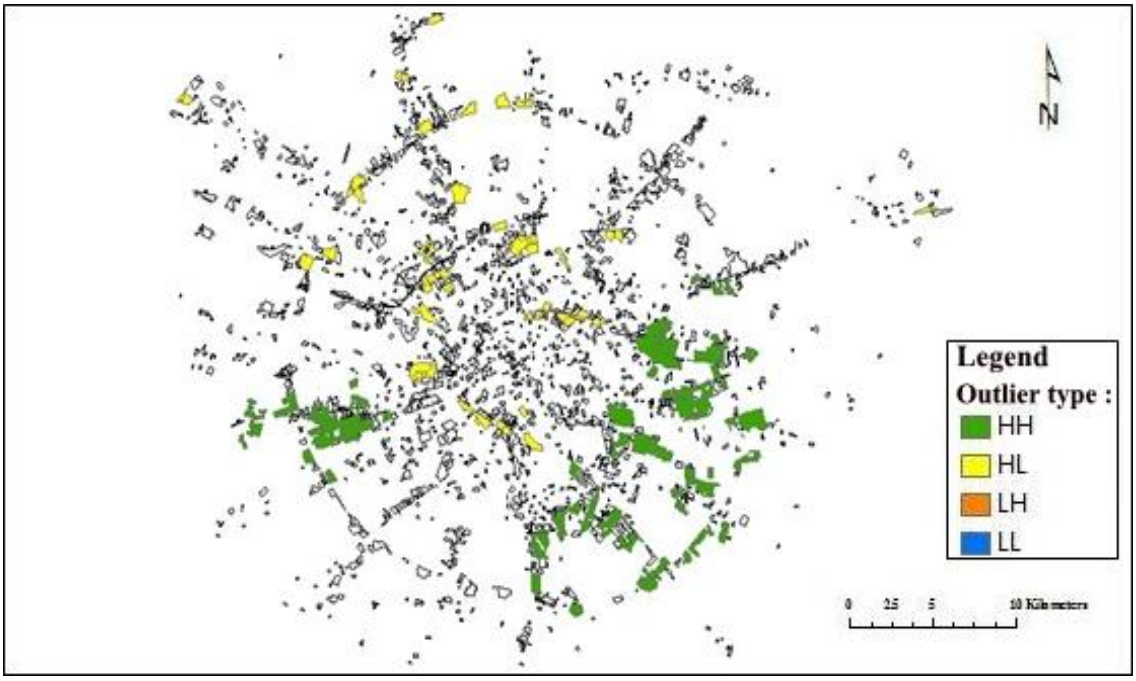


*Figure 5 : Bucharest – outliers detected using the method Zone of Indifference*

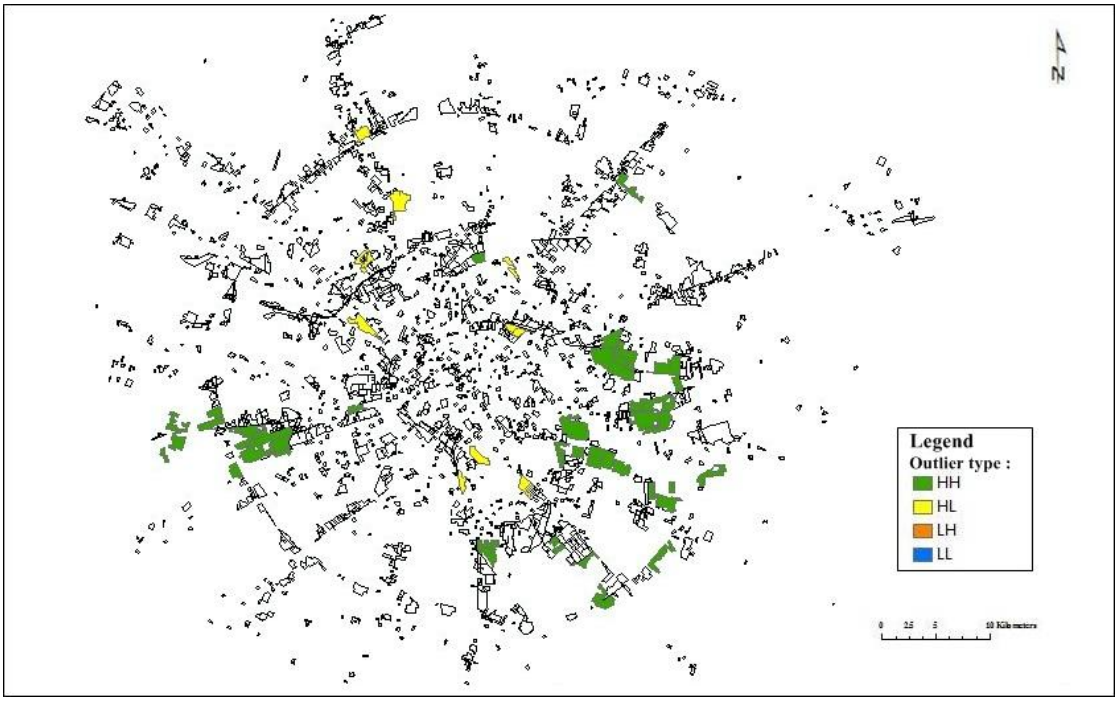


*Figure 6 : Bucharest – outliers detected using the method Inverse Distance Squared*

## Conclusions

It can be concluded that the methods Fixed Distance Band and Zone of Indifference identify the biggest number of outliers, in the most of the cases the number identified being identical for the both methods, while the method Inverse Distance Squared identify the smallest number of outliers. The fourth method used, Inverse Distance, identifies following a logical deduction a number of outliers which is intermediary, but in the most of the times closed to the number identified by the first two methods mentioned above. One possible explanation for the fact that the method Inverse Distance Squared identifies the smallest number of atypical values can be deduced from the fact that it analyses the space using the premise that it is anisotropic, thereby heterogenous and perhaps closer to the reality of the terrain . Thus, we can wrongly believe that this is perhaps the best method to perform the control quality for a database, but before jumping to this conclusion it should be considered that an outlier which is obtained by applying several methods is more likely to be wrong than an outlier which is obtained by applying only one method, so it is imperative that we do not limit only to a single method when identifying atypical values.

Therefore, one can conclude that the problem of identifying suspicious values remains a promising research area, which now is not used at its maximum potential, but which may open new perspectives for a study or a research because these outliers can reveal an atypical spatial distribution pattern which can provide valuable information about a particular phenomenon. Also, it allows the optimization of the quality of the databases used for various studies, because eliminating the outliers can significantly improve the applicability of a research, since it provides a more consistent binder between the theory and the reality from the field.

## References:


1. Anselin, L.,1996.*"The Moran scatterplot as an ESDA tool to assess local instability in spatial association",* in FISCHER, Manfred, SCHOLTEN, Henk K., UNWIN, David, Spatial analytical perspectives on GIS, Taylor & Francis, London, pp. 111-125.
2. Dixon, W. J., 1950. *Analysis of extreme values.* Annals of Mathematical Statistics, 21, 488-506.
3. Hawkins, D.M., 1980. *Identification of outliers.* London: Chapman and Hall.
4. Jarrell, M. G., 1994. *A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers.* Research in the schools, 1, 49-58.
5. Osborne, Jason W. & Amy Overbay, 2004. *The power of outliers (and why researchers should always check for them).* Practical Assessment, Research & Evaluation, 9(6)
6. Planchon V., 2005. *Traitement des valeurs aberrantes : concepts actuels et tendances générales*
7. Rasmussen, J. L., 1988. *Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. Multivariate Behavioral Research, 23(2),* 189-202.
8. Stevens, J. P., 1984. *Outliers and influential data points in regression analysis.* Psychological Bulletin, 95, 334-344.
9. ESPON Report Database, 2011. *Spatial analysis for quality control*
10. GMES Urban Atlas, Datasets, EEA - *(http://www.eea.europa.eu/data-and-maps/data/urban-atlas)*