Knowledge Discovery Process for Detection of Spatial Outliers

Giovanni Daián Rottoli[1,2,3][0000−0002−7623−2591], Hernán Merlino[3], and Ramón García-Martínez[3†]

[1] PhD Program on Computer Sciences. Universidad Nacional de La Plata. Argentina
[2] PhD Scholarship Program to Reinforce R+D+I Areas.
Universidad Tecnológica Nacional. Argentina.
[3] Information Systems Research Group. National University of Lanús. Argentina.
rottolig@frcu.utn.edu.ar, hmerlino@unla.edu.ar

Abstract. Detection of spatial outliers is a spatial data mining task aimed at discovering data observations that differ from other data observations within its spatial neighborhood. Some considerations that depend on the problem domain and data characteristics have to be taken into account for the selection of the data mining algorithms to be used in each data mining project. This massive amount of possible algorithm combinations makes it necessary to design a knowledge discovery process for detection of local spatial outliers in order to perform this activity in a standardized way. This work provides a proposal for this knowledge discovery process based on the Knowledge Discovery in Database process (KDD) and a proof of concept of this design using real world data.

Keywords: Spatial Outliers, Local Outliers, Spatial Data Mining, Knowledge Discovery Process, Spatial Clustering.

1 Introduction

Spatial outlier discovery is a knowledge discovery and data mining trend that has been used for many applications in fields such as climatology, geology, medicine, ecology and chemistry, among others [1–3].

Given a spatially referenced dataset $SD = \{d_i\}, i = 1..|SD|$, with each spatial object $d_i = [s_{i1}, s_{i2}, \ldots, s_{im}, v_{i1}, v_{i2}, \ldots, v_{in}]$, with spatial attribues $s_{ij} \in d_i, j = 1..m$ and non-spatial attribues $v_{ik} \in d_i, k = 1..n$, a spatial outlier is defined as an observation $d \in SD$, whose non-spatial attributes values differ from the non-spatial attributes values of its spatial neighbors [4]. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute "house age". Because of this, spatial outliers are considered *local outliers* [5].

Many methods have been developed for spatial outlier mining, such as Local Outlier Factor (LOF), Spatial Local Outlier Measure (SLOM), Novel Local Outlier Detection Method (NLOD), etc. [6–13], but some considerations should be taken into account.

First, according to Chandola et al. (2009) [14] and Liu et al. (2017) [15], it is a challenge nowadays to select the way in which neighborhoods are defined for local outlier detection without previous knowledge about the data domain. The contextual dependency of spatially referenced data causes that a spatial object may or may not be considered an outlier, according to the way in which neighborhoods were built. Then, a non-spatial attribute may vary in the space without it representing an abnormality. According to the given example, in a metropolitan area, the age of the buildings decreases gradually, with oldest ones located in the historical center. Lastly, a certain neighborhood may represent a spatial abnormality in relation to other neighborhoods, without spatial points being outliers.

Some of the methods proposed in the bibliography allow us to deal with some of the aforementioned considerations, but this shows the need to have several methods for outlier detection specialized in particular cases that can be varied to approach different situations, depending on the problem domain.

In this work, we propose a knowledge discovery process for detection of spatial outliers in order to have a standard procedure to perform this activity, regardless of the problem domain, the characteristics of the data and therefore, the data mining methods to be used.

This paper is organized as follows. Section 2 contains the knowledge discovery process for detection of spatial outliers. Section 3 shows how the process works using real-world data. Finally, conclusions and futures lines of work are presented in section 4.

2 Knowledge Discovery Process for Detection of Spatial Outliers

In this paper we propose a knowledge discovery process for detection of spatial outliers based on the Knowledge Discovery on Database process (KDD) [16], designed as a pipeline with well defined and separate activities to be able to use

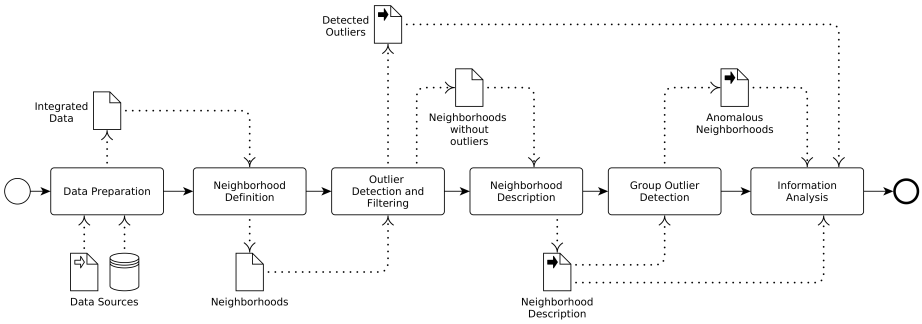

Fig. 1. Knowledge discovery process for detection of local spatial outliers, in Business Process Model Notation (BPMN)

Table 1. Integrated data format

Id	Spatial Attributes	Non-Spatial Attributes
id_1	s_{1j}	v_{1j}
id_2	s_{2j}	v_{2j}
$\vdots$	$\vdots$	$\vdots$
id_n	s_{nj}	v_{nj}

different data mining approaches for each of these tasks, in order to adapt the process to the particular data mining problem (Fig. 1).

This process includes steps for data preparation, neighborhood definition and description, with a proposal to be used in case of not having a predefined method to perform this activity; outlier detection; group outlier detection; and information analysis. All steps of the process are described in the following subsections.

2.1 Data Preparation

The first step for outlier detection is to integrate all the data, obtained from different sources into one single record of spatial objects consisting of spatial attributes, relative to the space, such as latitude and longitude, and non-spatial attributes, relevant to the problem domain (Table 1).

Each data source usually has its own format for data representation. For this reason all the sources must be integrated, cleaning errors, dealing with inconsistencies, and normalizing the attributes values if necessary. Also, it can be useful to use attribute selection algorithms such as Boruta [17] on non-spatial attributes if the number of columns is too large. Data preparation is a well-known activity of the KDD process [16].

2.2 Neighborhood Definition

After the construction of a unique data table, it is necessary to select a criterion for the construction of neighborhoods: as mentioned before, depending on the way in which neighborhoods are defined, some spatial points may or may not be considered outliers.

This criterion is related to the problem domain; in some cases, there might be a predefined method to consider two or more points as neighbors, and in other cases the business intelligence and data analysts might ignore it. Because of this, we propose the use of unsupervised learning using spatial clustering techniques to perform this activity if the criterion to group the data is unknown.

In previous works [12, 15], spatial clustering techniques were used for outlier detection considering the points that do not belong to any cluster (or single-point clusters) as anomalous. In this work, on the other hand, different clustering algorithms can be used for definition of the regions that presents a homogeneous

Table 2. Integrated Data format with neighborhoods

Id	Spatial Attributes	Non-Spatial Attributes	Neighborhood
id_1	s_{1j}	v_{1j}	N_1
id_2	s_{2j}	v_{2j}	N_2
$\vdots$	$\vdots$	$\vdots$ $\quad \vdots$	
id_n	s_{nj}	v_{nj}	N_l

behavior in relation to the non-spatial attributes of the spatial data. The points that differ from the normal behavior of its neighborhood will be considered outliers.

This work proposes the use of REDCAP algorithms for spatial clustering. This graph-based algorithm family is aimed at discovering clusters of arbitrary shape with all its elements contiguous to each other. This contiguity criterion, as mentioned before, depends on the problem domain [18–20].

As result of this activity a new column with the neighborhood to which each spatial object belongs is added to the original data table (Table 2).

2.3 Outlier Detection and Filtering

The spatial data with the spatial neighborhoods is used as input in order to discover spatial objects that do not have a normal behavior in relation to other spatial objects in its neighborhood. Those abnormal points will be considered outliers.

To perform this analysis a measure of outlierness must be calculated for each spatial point in each neighborhood using its non-spatial attributes to discover the deviations in each cluster, considering that these attributes were used to group similar spatial points. Different algorithms can be used, such as LOF, SLOM, different distance measures like Euclidean or Mahalanobis, among others.

This work proposes the use of Mahalanobis Distance [21, 22] to calculate how much do the non-spatial values of a spatial object differ from the median values of their neighbors. The reason for using this method is the fact that this distance takes into account the correlations of the data set, but a few different methods can be used depending on the problem domain. For example, LOF-based algorithm can be used to consider density between points, or even more complex methods can be applied, such as the ones proposed by Kuna et al. (2012) [23].

After calculating the outlierness value, the spatial points with a value higher than a user-specified threshold must be filtered and two different tables will be generated: one with the anomalous points and one with the non-anomalous points.

2.4 Neighborhood Description

In this step, after filtering the outliers from the database, it is necessary to describe the normal behavior of each neighborhood in order to use it later in the analysis step to find out the characteristics of the abnormal data objects.

Unsupervised learning discovers implicit patterns from data, but they must be described according to the values of its non-spatial attributes to provide an easy understanding of the process results to analysts.

Several methods can be used to perform this activity and all of them can be complemented. For example, the application of rule-based classification algorithms such as C4.5 [24, 25] or CART [26, 27] is a good tool to find out the decision rules that describe each neighborhood [20], as well as classical statistical measures such as mean and standard deviation of each non-spatial data attribute. The selection of the tools and algorithms also depends on the problem domain and the data characteristics.

The result of this step is a description of each spatial neighborhood which will be used later to detect outlier clusters.

2.5 Group Outlier Detection

As mentioned before, some spatial clusters could have very different behavior to its spatial neighbors. No progression can be noticed between adjacent clusters. An example of this is shown in Figure 2.

Consider this image as a tree plantation where each pattern corresponds to a spatial cluster or spatial area with a different kind of tree: Eucalyptus, Pinus and Teak. The description of each cluster will be based on the spatial object attribute Species, but, if we consider another attribute such as the rotation age of each kind of tree, we can know that the Teak tree cluster is anomalous in relation to its neighbors because of its old age (63.3 years in average) compared to its neighbors: an average of 8.2 years in the case of eucalyptus, and 20.16 years for the pinus species [28].

Because of this, it is possible to model each spatial cluster as spatial objects with its non-spatial attributes equal to the description data obtained in the last

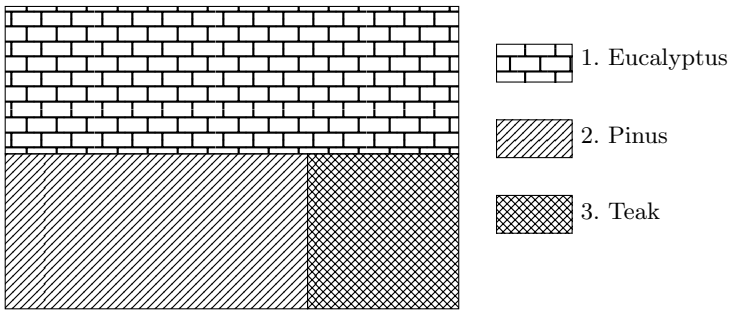

Fig. 2. Example of tree plantation distribution.

Table 3. Description table for the data cluster examples (may not be correct).

Id	Species	Rotation Length Avg.	Rotation Length S.D.	...
1	Eucalyptus	8.2	0.83	...
2	Pinus	20.16	3.43	...
3	Teak	63.3	5.77	...

step and using the adjacency between them as neighboring criteria. The example shown above is modeled in Table 3.

In this spatial clusters model, any of the data outlierness values mentioned in section 2.3 can be used to assess whether any cluster deviates too much from its neighbors. In the same way as in the previous cases, the model depends on the problem domain and the analyst should select the methods they deem more appropriate.

2.6 Information Analysis

At this point of the process, many information resources were created. These resources are (i) the list of the spatial data objects marked as outliers, (ii) the description of the normal behavior of the neighborhoods, and (iii) the oulierness level of each neighborhood. This information has to be analyzed to generate useful knowledge for decision makers.

The neighborhood descriptions can be used in order to analyze each of the outliers detected: if you know the normal behavior of the neighborhood, the abnormal values in the data objects can be found. Also, these descriptions can be used to find out the differences between the neighborhoods detected as outliers and the normal ones. All these activities can be automatized in order to speed up the analysis.

3 Proof of Concept

In this section a real dataset and simple data mining algorithms were used to show how the proposed process works in a real scenario, not with an emphasis on the results of the outlier search, but on the operation of the proposed process using real data.

The real dataset used in this proof of concept corresponds to county population data from the United States of America [29]. From this data file, only counties from 5 states were used in order to reduce the dataset size: Illinois, Indiana, Kentucky, Missouri and Tennessee. Also, only four non-spatial attributes were selected from the original data file, and two new spatial attributes were generated from geographical data: 'POPESTIMATE2016', 7/1/2016 total resident population estimate, renamed as 'Pop'; 'BIRTHS2016', Births in period 7/1/2015 to 6/30/2016, renamed as 'Births; 'DEATHS2016', Deaths in period

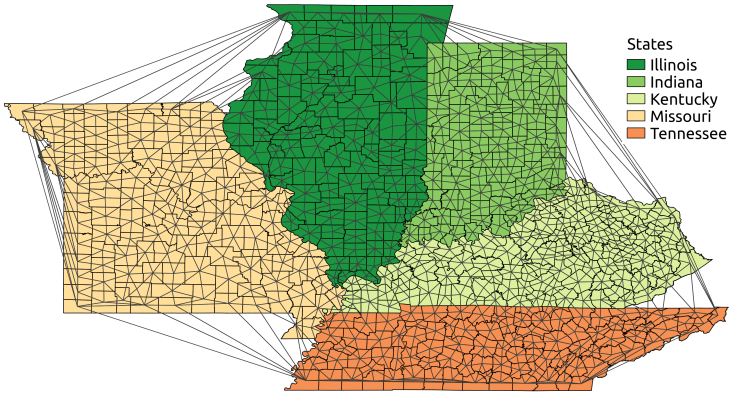

Fig. 3. Delaunay triangulation as contiguity criterion between counties centroids.

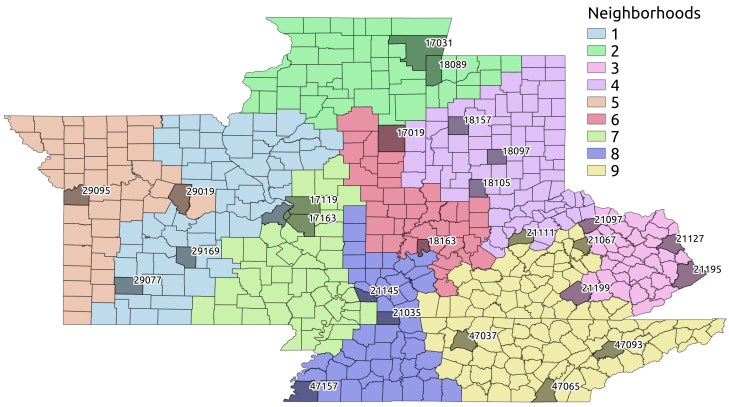

Fig. 4. Detected Neighborhoods in USA Counties. Shaded counties were detected as outliers.

7/1/2015 to 6/30/2016, renamed as 'Deaths'; 'INTERNATIONALMIG2016', Net international migration in period 7/1/2015 to 6/30/2016, renamed as 'Mig'; 'Lat', generated attribute with the Latitude of the county centroid, and 'Long', generated attribute with the longitude of the county centroid. Records with empty fields were also removed, resulting in 524 rows.

This integrated dataset was used as input for REDCAP first-order single-linkage regionalization algorithm [18] considering two counties as contiguous if their centroids are linked with an edge in the county centroid's Delaunay triangulation (Fig. 3). This algorithm is not the most efficient from the REDCAP family, but it was chosen because of its simplicity to carry out the proof of concept. This activity resulted in nine clusters or neighborhoods (Fig. 4).It should be noted that the Haversine formula was used to calculate the distance between spatial points.

Table 4. Outliers discovered using the proposed process, ordered by neighborhood and GeoID

	Non-Spatial Attributes				Spatial Attributes		
GeoID	Pop	Births	Deaths	Mig	Long	Lat	Neighborhood
29077	288690	3590	2753	341	-93.34199	37.25805	1
29169	52654	780	271	233	-92.20766	37.82463	1
29189	998581	11591	10027	2186	-90.44341	38.64068	1
17031	5203499	68049	42297	18434	-87.64616	41.8954	2
18089	485846	5918	4908	324	-87.37636	41.4722	2
21097	18646	215	205	21	-84.33139	38.44181	3
21127	15863	200	210	21	-82.73475	38.06788	3
21195	60555	623	831	11	-82.39587	37.46903	3
21199	63956	716	795	38	-84.57718	37.10393	3
18105	145496	1305	894	1019	-86.52314	39.16092	4
18097	941229	14433	8015	3033	-86.13847	39.78171	4
18157	188059	2356	1138	1716	-86.8941	40.38862	4
29019	176594	2119	1064	698	-92.30966	38.99062	5
29095	691801	9423	6473	895	-94.34609	39.0085	5
17019	208419	2413	1345	1798	-88.19919	40.14008	6
18163	181721	2188	1887	186	-87.58578	38.02514	6
17119	265759	3032	2814	148	-89.90517	38.82985	7
17163	262759	3282	2678	127	-89.92841	38.4703	7
29510	311404	4547	3070	981	-90.24512	38.63581	7
21035	38437	407	408	119	-88.2722	36.6211	8
21145	65162	728	848	9	-88.71272	37.05408	8
47157	934603	13448	8326	1258	-89.8956	35.184	8
21067	318449	4107	2231	1350	-84.45873	38.04233	9
21111	765352	9918	7513	2460	-85.65916	38.18719	9
47037	684410	10438	5461	3091	-86.78482	36.16944	9
47065	357738	4227	3529	527	-85.16479	35.18086	9
47093	456132	5285	4321	598	-83.93709	35.99322	9

After neighborhood definition, the Mahalanobis distance was calculated for each object as an outlierness measure using only its non-spatial attributes , as shown in Formula 1, where $\boldsymbol{s_j}$ is the vector of non-spatial values of the evaluated point, $\boldsymbol{\mu}$ is the vector of non-spatial mean values from the neighborhood, and C is the covariance matrix.

$$\sqrt{(\boldsymbol{s_j} - \boldsymbol{\mu})' \mathbf{C}^{-1} (\boldsymbol{s_j} - \boldsymbol{\mu}))} \tag{1}$$

Afterwards, the data records with a distance higher than one standard deviation of the mean distance of the other spatial objects in the same neighborhood were filtered out as outliers(Fig. 4). As a result of this activity, two datasets were available: one with only outlier counties (Table 4), and one with the original dataset without the outliers.

Table 5. Neighborhood description using statistical measures with the distance between them. Outlier clusters are shown in bold.

	Mean				Standard Deviation				
N	Pop	Births	Deaths	Mig	Pop	Births	Deaths	Mig	M. Dist
1	33754.67	385.57	331.81	29.22	56192.63	653.11	441.8	85.29	6.93
2	136820.51	1601.74	1092.69	219.86	206767.29	2403.60	1390.69	527.26	7.07
3	**18862.3**	**223.78**	**234**	**3.25**	**12602.97**	**153.18**	**151.7**	**4.71**	**2.52**
4	51439.08	624.69	473.85	56.08	59744.4	787.73	466.96	132.15	5.86
5	30888.16	377.32	291.58	32.88	43050.14	553.7	332.66	67.98	6.78
6	29718.71	351.25	305.15	29.3	32312.55	387.77	276.28	99.9	6.93
7	28076.34	316.02	314.86	17.15	34015.47	387.5	314.68	42.58	4.51
8	**25599.76**	**293.5**	**304.91**	**10.91**	**18538.67**	**234.24**	**203.67**	**17.11**	**2.61**
9	45742.36	538.35	455.56	50.61	49057.77	640.15	381.33	112.18	4.74

The dataset without outliers was described using simple statistical measures: for each neighborhood, the mean and standard deviation of each non-spatial attribute were calculated, yielding the values shown in Table 5. Considering each of these descriptions as a spatial object and considering all clusters as neighbors, the Mahalanobis distance was also calculated, providing the distance between each description in order to discover anomalous clusters. Neighborhoods with a distance that differs from the mean distance by more than 1 standard deviation are considered outliers .

Using all the information obtained, it is possible to acquire knowledge about the anomalous data behavior using the filtered outliers and the neighborhood descriptions: for example, counties with GEOId 29077 and 29183 are two of the outliers discovered in neighborhood 1 (Table 4), these outliers have values considerably greater than the values of the description of the neighborhood. On the other hand, it must be highlighted that the other outlier in this neighborhood also has bigger attribute values than its neighborhood description, with the exception of the number of deaths, which is smaller. This knowledge could be of interest for the business intelligence.

Additionally, using the neighborhood description data with the Mahalanobis distances, it can be seen that outlier clusters 3 and 8 are mainly characterized by a small number of migrations compared to the other clusters.

Also, combining this information it can be noticed that the anomalous neighborhood 8 has three outlier counties, but only one of them has a number of migrations in the same order than its description, in spite of having higher values for its other attributes. These characteristics can be valuable for decision making.

Lastly, a look at the outlier distribution on the map may be worthwhile (Fig. 4): the majority of the outliers discovered are located at the borders of the generated regions.A refinement of the definition of these neighborhoods can be performed by going back to this stage of the proposed process and repeating the

detection of outliers, in case they are caused by an error of the regionalization algorithm.

4 Conclusion

This work describes a knowledge discovery process for detection of spatial outliers. This process is designed as a pipeline based on the KDD Process with properly divided activities that include data preparation, neighborhood definition and description, outlier detection and filtering, group outlier detection, and information analysis using the data obtained in each step. These activities are independent from data mining algorithms, allowing data scientists to use the algorithms they deem appropriate to solve the problems related to the problem domain or the data they are working with.

A proof of concept of this knowledge discovery process was provided using real world data, basic algorithms for neighborhood definition, and statistical measures for outlier detection and outliers and neighborhood description, in order to show how the process works in a real setting.

An evaluation for different data mining algorithms must be carried out to find a good combination thereof for well-known problems, in order to have a battery of configurations ready to be used as a starting point in spatial outlier analysis. Also, it is possible to generate a heuristic approach for data mining algorithm selection using problem domain metadata. Lastly, the use of this process into an iterative methodology such as CRISP-DM [30] or MoProPEI [31] must be assessed.

Acknowledgments

The research presented in this paper was partially funded by the PhD Scholarship Program to reinforce R&D&I areas (2016-2020) of the Universidad Tecnológica Nacional, Research Project 80020160400001LA of National University of Lanús, and PIO CONICET-UNLa 22420160100032CO of National Research Council of Science and Technology (CONICET), Argentina. The authors also want to extend their gratitude to Kevin-Mark Bozell Poudereux, for proofreading the translation, and the anonymous reviewers of this work for their valuable comments and suggestions.

References

1. Araki, Shin, et al. "Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan". Atmospheric Environment 153 (2017): 83-93.
2. Bakon, Matus, et al. "A Data Mining Approach for Multivariate Outlier Detection in Postprocessing of Multitemporal InSAR Results". IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2017).

3. Bobbia, Michel, et al. "Spatial outlier detection in the air quality monitoring network of Normandy (France)". GRASPA WORKING PAPERS (2014).
4. Deepak, P. "Anomaly Detection for Data with Spatial Attributes". Unsupervised Learning Algorithms. Springer International Publishing. (2016): 1-32.
5. Shekhar, Shashi, Chang-Tien Lu, and Pusheng Zhang. "A unified approach to detecting spatial outliers". GeoInformatica 7.2 (2003): 139-166.
6. Breunig, Markus M., et al. "LOF: identifying density-based local outliers". ACM sigmod record. 29.2. ACM. (2000).
7. Chawla, Sanjay, and Pei Sun. "SLOM: a new measure for local spatial outliers". Knowledge and Information Systems 9.4 (2006): 412-429.
8. Schubert, Erich, Michael Weiler, and Arthur Zimek. "Outlier Detection and Trend Detection: Two Sides of the Same Coin". Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE.(2015).
9. Kamble, Bharati, and Kanchan Doke. "Outlier Detection Approaches in Data Mining". International Research Journal of Engineering and Technology (IRJET). 4.3 (2017): 634-638.
10. Ernst, Marie, and Gentiane Haesbroeck. "Comparison of local outlier detection techniques in spatial multivariate data". Data Mining and Knowledge Discovery 31.2 (2017): 371-399.
11. Tang, Bo, and Haibo He. "A local density-based approach for outlier detection". Neurocomputing 241 (2017): 171-180.
12. Du, Haizhou, et al. "Novel clustering-based approach for Local Outlier Detection". Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on. IEEE. (2016).
13. Liu, Xutong, Chang-Tien Lu, and Feng Chen. "Spatial outlier detection: Random walk based approaches". Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM.(2010).
14. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". ACM computing surveys (CSUR) 41.3 (2009): 15.
15. Liu, Qi, et al. "Unsupervised detection of contextual anomaly in remotely sensed data". Remote Sensing of Environment (2017).
16. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data". Communications of the ACM 39.11 (1996): 27-34.
17. Kursa, Miron B., Aleksander Jankowski, and Witold R. Rudnicki. "Borutaa system for feature selection". Fundamenta Informaticae 101.4 (2010): 271-285.
18. Guo, Diansheng. "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)". International Journal of Geographical Information Science 22.7 (2008): 801-823.
19. Mennis, Jeremy, and Diansheng Guo. "Spatial data mining and geographic knowledge discoveryAn introduction". Computers, Environment and Urban Systems 33.6 (2009): 403-408.
20. Rottoli, Giovanni Daián, Hernan Merlino, and Ramón García-Martínez. Knowledge Discovery Process for Description of Spatially Referenced Clusters. International Conference on Software Engineering & Knowledge Engineering. Ed. USA KSI Research Inc. and Knowledge Systems Institute. (2017): 410415. Web. DOI: 10.18293/SEKE2017-013
21. De Maesschalck, Roy, Delphine Jouan-Rimbaud, and Désiré L. Massart. "The mahalanobis distance". Chemometrics and intelligent laboratory systems 50.1 (2000): 1-18.

22. Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies". Artificial intelligence review 22.2 (2004): 85-126.
23. Kuna, Horacio, Ramón García-Martínez, and Francisco Villatoro. "Automatic Outliers Fields Detection in Databases". Journal of Modelling and Simulation of Systems 3.1 (2012): 14-20.
24. Quinlan, J. Ross. "Improved use of continuous attributes in C4. 5". Journal of artificial intelligence research 4 (1996): 77-90.
25. Quinlan, J. Ross. "C4. 5: programs for machine learning". Elsevier.(2014).
26. Breiman, Leo, et al. "Classification and regression trees". CRC press. (1984).
27. Bel, Liliane, et al. "CART algorithm for spatial data: Application to environmental and ecological data". Computational Statistics & Data Analysis 53.8 (2009): 3082-3093.
28. Ugalde Luis and Osvaldo Pérez. Table 1. "Productivity and rotation lengths for main forest plantation trees in selected tropical countries. Mean Annual Volume Increment of Selected Industrial Forest Plantation Species". Forest Plantations Thematic Papers. Forestry Department of Food and Agriculture Organization of the United Nations. (2001).
29. United States Census Bureau. "Population, population change and estimated components of population change: April 1, 2010 to July 1, 2016". (CO-EST2016-alldata). County Population Totals Datasets: 2010-2016. On-Line: https://www.census.gov/data/datasets/2016/demo/popest/-counties-total.html (October 17, 2017)
30. Wirth, Rdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. (2000).
31. Martins, Sebastian, Patricia Pesado, and Ramón García-Martínez. "Intelligent systems in modeling phase of information mining development process". International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer International Publishing. (2016). DOI:10.1007/978-3-319-42007-3_1