Knowledge Discovery Process for Description of Spatially Referenced Clusters

Giovanni Daián Rottoli [1,2,3], Hernán Merlino [3], Ramón García-Martinez [3,4 †]
[1] PhD Program on Computer Sciences. National University of La Plata. Argentina.
[2] PhD Scholarship Program to Reinforce R+D+I Areas. Technological National University. Argentina
[3] Information Systems Research Group. National University of Lanús. Argentina
[4] Scientific Research Commission - CIC Bs As, Argentina. Argentina
gd.rottoli@gmail.con, hmerlino@gmail.com

Abstract— **Spatial clustering is an important field of spatial data mining and knowledge discovery that serves to partition a spatial data set to obtain disjoint subsets with spatial elements that are similar to each other. Existing algorithms can be used to perform three types of cluster analyses, including clustering of spatial points, regionalization and point pattern analysis. However, all these existing methods do not provide a description of the discovered spatial clusters, which is useful for decision making in many different fields. This work proposes a knowledge discovery process for the description of spatially referenced clusters that uses decision tree learning algorithms. Two proofs of concept of the proposed process using different spatial clustering algorithm on real data are also provided.**

Keywords- Knowledge Discovery Process; Spatial Clustering; Regionalization; Decision Tree Learning; Spatial Data Mining.

I. INTRODUCTION

Clustering is an information mining and knowledge discovery task that can be used to partition a spatial data set to obtain disjoint subsets whose elements are similar to each other [1, 2].

There are three ways to consider spatial information in clustering analysis: (i) spatial clustering to find disjoint groups of spatial points using spatial attributes, non-spatial attributes or both; (ii) regionalization, adding a spatially contiguity constraint between spatial objects to spatial clustering, and (iii) point pattern analysis, to detect unusual concentrations of spatial points in some regions of the space [3].

This variety adds a level of complexity to the data mining task [3] and, because of that, new algorithms and methods for spatial clustering have been developed in recent years: REDCAP algorithms for regionalization; DBSCAN, NSCABDT, ASCDT, among others, for clustering of spatial points; and DBSCAN and RDBC for point patterns analyses, to name a few [3, 5-16].

However, these algorithms just make use of different strategies to generate groups or clusters between the different spatially referenced objects, but none of them serves to obtain, in a systematic way, the description of the automatically generated clusters in order to know which criteria were used to realize this activity.

For this reason, a knowledge discovery process, defined as a group of logically related tasks that form a set of information with a degree of value for the organization obtains knowledge pieces that generalize the previous information [1, 17, 18], is designed to generate decision rules on clustering results regardless of the selected approach to generate spatial clusters.

The remainder of this paper is organized as follows: Section 2 describes the problem derived from the analysis of the state-of-the-art. In Section 3, we propose a Knowledge Discovery Process to solve the problem described. In Section 4, two proofs of concept are presented using real data. Finally, conclusions derived from the research are outlined in Section 5.

II. PROBLEM DEFINITION

Many algorithms and methods were developed to discover spatially referenced clusters in any of its forms: spatial clusters, regions or point patterns. These algorithms make use of different heuristics and techniques to separate the spatial objects more similar to each other, according to a specific similarity function. As result of this task, for each spatial object, the cluster to which it belongs is specified. However, as mentioned before, these algorithms do not allow describing the automatically discovered clusters according to the attributes chosen for that activity.

This issue affects the decision makers that use this kind of data mining algorithms. When spatial clustering algorithms are used and because clusters are not relevant by themselves, the results obtained have to be analyzed and visualized to retrieve relevant information for decision making in a certain problem domain.

If there is no systematic way to conduct this activity, decision-makers should make an extra effort to interpret the results or should be necessary to spent more resources using experts to do this task previously.

For this reason, it is interesting to design a knowledge discovery process, such as those proposed in [1], which can be used to obtain partitions of the information mass in a systematic way and then describe each partition or cluster according to the values of the attributes of the data that belong to each of them.

DOI reference number: 10.18293/SEKE2017-013

III. PROPOSED SOLUTION

García-Martinez *et al.* (2013) proposed a knowledge discovery process for Group-Membership Rules to identify the conditions of membership to each of the classes of an unknown partition *a priori*, but existing in the available information bases of the problem domain. This process makes use of Top Down Induction of Decision Tree algorithms (TDIDT) on the result of applying a clustering algorithm, e.g. Self-Organizing Maps (SOM), on transactional data to determine the conditions to belong to a group. Based on this previous work, a knowledge discovery process for description of spatially referenced clusters is proposed in this paper following the same concept.

This process, as mentioned before, serves to find spatial clusters in any of its forms (i.e., regionalization, clustering of spatial points and point pattern analysis), and find the rules that describe the characteristics of each of them, based on the data attributes selected to be used for clustering. Figure 1 shows the proposed process using Business Process Model Notation [19].

The process, as can be seen in Figure 1, takes a set of spatially referenced data as input represented in different formats such as, *inter alia*, plain text, databases, tables and geographic information system maps. These data are integrated to a table comprised of the object identifier, spatial attributes (e.g. object location), and non-spatial attributes such as object properties according to the problem domain.

The integrated data are used for cluster discovery process [1]. For this purpose, it is necessary to select a type of spatial cluster among the aforementioned types: regionalization, clustering of spatial points or point patterns analysis, according to the problem domain, and choose appropriate algorithms in each case.

This paper proposes the use of any of the REDCAP algorithms for region generation, because of their benefits over other regionalization algorithms [6]. In this case, it is necessary to provide contiguity constraints between the spatial objects as algorithm input.

On the other hand, both in the case of clustering of spatially referenced objects and point pattern analyses, it is suggested to use density-based algorithms[11,14], such as DBSCAN-like algorithms [11, 15, 16], DENCLUE [20], ASCDT [9] o DBSC [10], because of the same reason mentioned above.

In each case, the input attributes depend on the selected algorithm. After the clustering algorithm execution, it is necessary to create a new table comprised of the integrated information with a new attribute in which each row registers the cluster to which the spatial object belongs.

In the last step, a Decision Tree Learning algorithm, such as C4.5 [21] or Random Forest [22-24], to name a few, is used to generate the rules that describe the characteristics of each cluster. For this purpose, it is necessary to identify the input attributes and the target attribute beforehand: input attributes will be the non-spatial data attributes, and target attribute will correspond to the attribute added in the last step, which has information about the automatically generated clusters. The parameters for the decision tree learning algorithms have to be selected depending on the particularities of the data.

As a result of the proposed knowledge discovery process, a set of rules that describes the automatically generated clusters is provided.

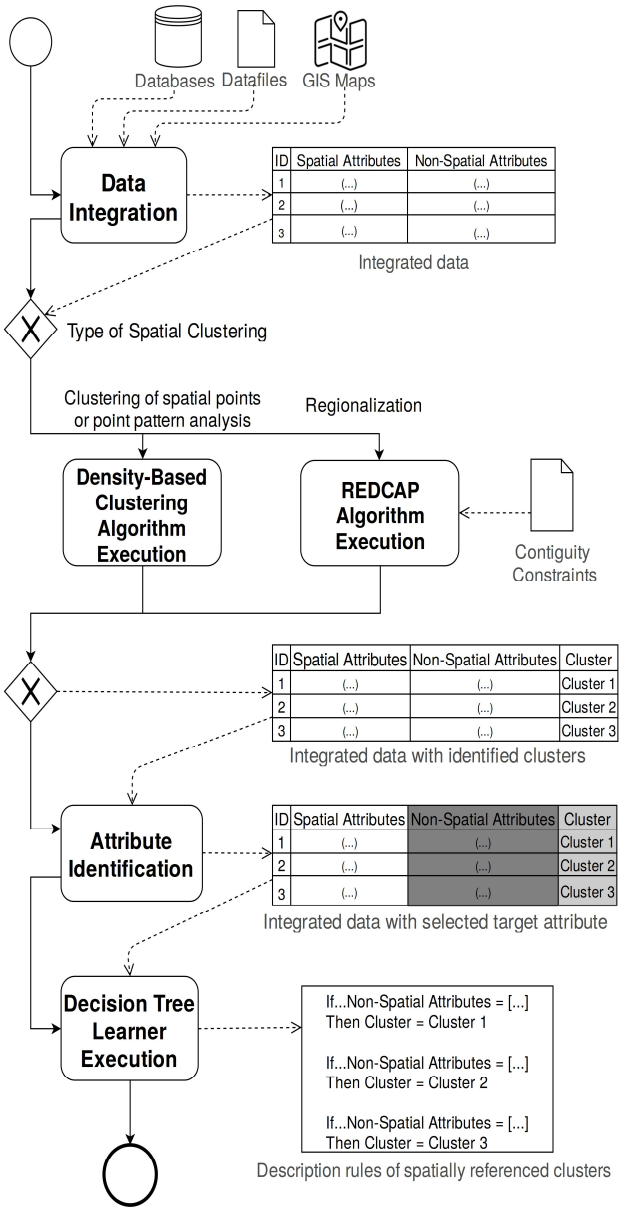

Figure 1. Knowledge discovery process for description of spatially referenced clusters

IV. PROOFS OF CONCEPT

This section includes two proofs of concept of the proposed process using real datasets obtained from different sources. In subsection 4.1 regionalization algorithms are used on demographic data, and in subsection 4.2 density-based clustering algorithms are applied to meteorological stations data.

A. *Demographic Data Regionalization*

In this proof of concept of the proposed knowledge discovery process, demographic data of 100 cities from the state of Iowa (IA) of the Unites States of America, updated on October 4th, 2016 and obtained from [26] were used for regionalization.

In the first step of the process, the selected data were integrated and normalized into a set of attributes, as shown in Table 1, being all of them numeric attributes.

Then, a regionalization algorithm from the REDCAP family of regionalization algorithms was selected: the First Order-SLK algorithms. This algorithm is not the REDCAP algorithm with a better behavior [6], but the simplest among them and, because of that, it was chosen to illustrate how the proposed process behaves. In this case, Delaunay triangulation was selected as contiguity criterion: two cities are contiguous if they are connected by an edge in the Delaunay triangulation of all the cities, using the Euclidean distance between them [25].

After the application of this algorithm selecting 6 as the number of desired regions to be obtained, the regions shown in Figure 2 using Voronoi diagrams [27] with each city as generator were discovered. Then, those regions were integrated in a single record with the original data.

Later, the Top Down Induction of Decision Tree algorithm C4.5 [20] was selected to generate the rules that describe the discovered regions, using Tanagra's implementation [28] and selecting the discovered region column in the integrated data as target attribute, and the demographic attributes "Population", "Households", "Median_income" and "Land_area" as input attributes.

As a result, we obtained 13 rules. For each region there is one or more rules than describe it. If the region is described with more than a single rule, it means that the region contains non-similar elements. Figure 3 shows the rules of the regions described with a single rule. On the other hand, the rules of the regions described with more than a single rule are shown in Figure 4.

As shown in Figure 3, regions 1, 2 and 3 are described with a single rule. Region 1 has more than 30 households, a population greater than 107 but lower than 2394, median incomes lower than USD72500 and a land area between 6236971.5 and 21839152 square meters. Region 2 has a very simple description: if the land area of the city is greater than 21839152 square meters, the city belongs to region 2. Finally, the rule that describes region 3 has a low confidence; it means that the cities that belong to it are not very similar, only 42.86% of the cities correspond to the description. The remaining 57.14% are not similar to each other and, because of that, there is no other rule that describes it.

The regions described with many rules, such as regions 0, 4 and 5 (Figure 4), can be divided into new regions to obtain more homogeneous sub-regions. For instance, region 5 is described with 2 rules with over 60% confidence: the first rule refers to cities with a land area between 4113504 and 6236971.5 square meters, a population between 107 and 2394, a median income lower than USD72500 and over 30 households, while the second one refers to cities with less than 30 households and a land area lower than 21 839 152 square meters. If more regions are specified, it is possible to divide the regions described using many rules into many different and homogeneous regions. However, this depends on the heterogeneity function used in the regionalization algorithm, which in turn depends on the domain of the problem.

TABLE I. DESCRIPTION OF THE ATTRIBUTES OF THE DEMOGRAPHIC DATA USED IN THE FIRST PROOF OF CONCEPT

Attribute	Description
Id	City Identifier
Latitude	City coordinates
Longitude	
Population	The total population living within city limits, using the latest US Census 2014 Population Estimates
Households	The total number of households within city limits using the latest 5 year estimates from the American Community Survey.
Median_income	The average (median) household income for the record using the latest 5 year estimates from the American Community Survey (USD)
Land_area	The area of land covered by the city in sq. meters.

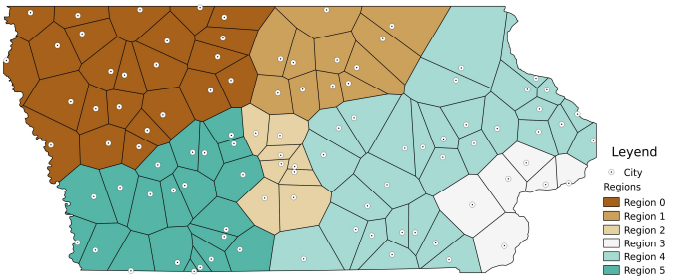

Figure 2. Regionalization of cities of the state of Iowa, USA

```
IF  "households" >= 30
AND "population" >= 107
AND "population" < 2394
AND "median_income" < 72 500
AND "land_area" < 21 839 152
AND "land_area" >= 6 236 971.5
THEN Region = 1  [Confidence = 66.67%]

IF "land_area" >= 21 839 152
THEN Region = 2  [Confidence = 83.33%]

IF "households" >= 30
AND "median_income" < 72 500
AND "land_area" < 1 011 569.5
AND "population" >= 245
THEN Region = 3  [Confidence = 42.86%]
```

Figure 3. Description rules of regions described with a single rule, obtained using the proposed process.

```
If "land_area" >= 1 011 569.5
And "land_area" < 1 427 927.0
And "median_income" < 72500.0
And "population" >= 106.5
And "households" >= 30
Then Region = 0 [Confidence = 85.71%]

If "population" < 245
And "population" >= 106.5
And "land_area" < 1 011 569.5
And "median_income" < 72 500
And "households" >= 30
Then Region = 0 [Confidence = 71.43%]

If "population" >= 522,5000
And "land_area" >= 2 579 852.5
And "land_area" < 4 113 504
And "median_income" < 72500,0000
And "households" >= 30,0000
Then Region = 0 [Confidence = 83.33%]

If "land_area" < 2 579 852.5
And "land_area" >= 1 760 551.5
And "median_income" < 72 500
And "population" >= 106.5
And "households" >= 30
Then Region = 0 [Confidence = 0,5556]

If "population" < 106
And "households" >= 30
And "land_area" < 21 839 152
Then Region = 4 [Confidence = 60%]

If "median_income" >= 72 500
And "population" >= 106.5
And "households" >= 30
And "land_area" < 21 839 152
Then Region = 4 [Confidence = 60%]

If "population" >= 2 394,5
And "land_area" >= 4 113 504
And "land_area" < 21 839 152
And "median_income" < 72 500
And "households" >= 30
Then Region = 4 [Confidence = 57.14%]

If "land_area" < 1 760 551.5
And "land_area" >= 1 427 927
And "median_income" < 72 500
And "population" >= 106.5
And "households" >= 30
Then Region = 4 [Confidence = 77.78%]

If "land_area" < 6 236 971.5
And "land_area" >= 4 113 504
And "population" < 2 394.5
And "population" >= 106.5
And "median_income" < 72 500
And "households" >= 30
Then Region = 5 [Confidence = 71.43%]

If "households" < 30
And "land_area" < 21 839 152
Then Region = 5 [Confidence = 64.29%]
```

Figure 4. Description rules of regions described with many rules, obtained using the proposed process.

B. Spatial Point Clustering

For this second proof of concept, data obtained from meteorological stations in Argentina on June 6, 2016 were used for this experiment [29]. The considered attributes of the mentioned data are shown in Table 2. The values were normalized.

TABLE II. DESCRIPTION OF THE ATTRIBUTES OF THE METEOROLOGICAL DATA USED IN THE SECOND PROOF OF CONCEPT

Attribute	Description
Lat	Meteorological Station Coordinates
Long	
TMin	Minimum temperature measured on June 6, 2016
TMax	Maximum temperature measured on June 6, 2016
TAv	Average temperature measured on June 6, 2016

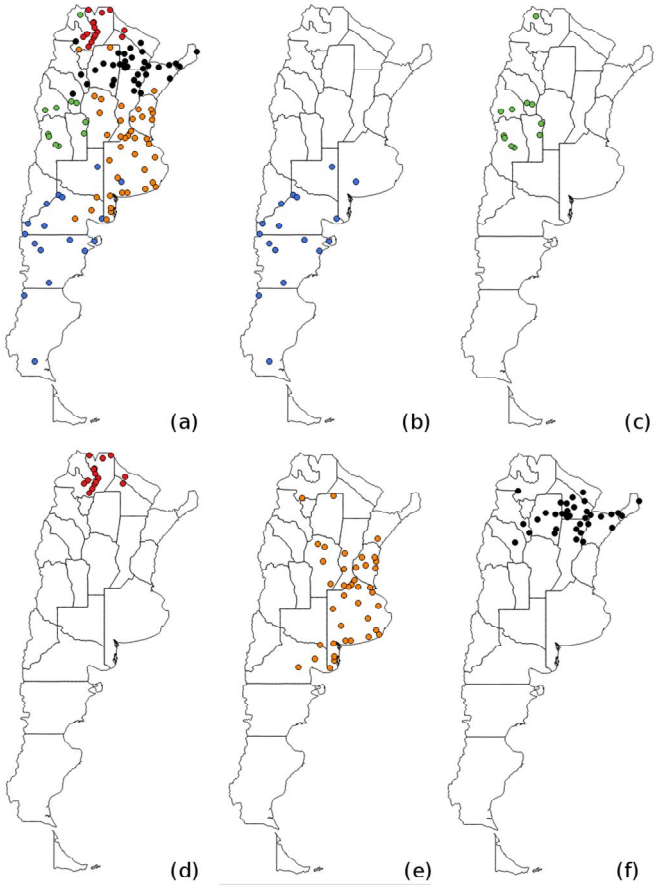

Figure 5. (a) Argentinean meteorological stations distribution with data on June 6, 2016. (b) Meteorological stations in cluster 0 (c) Meteorological stations in cluster 1. (d) Meteorological stations in cluster 2. (e) Meteorological stations in cluster 3. (f) Meteorological stations in cluster 4.

The algorithm selected for the proof of concept was DBSCAN [15], a density-based algorithm for clustering using the implementation available in the data mining software WEKA [30], resulting in 5 spatial clusters with the distribution shown in Figure 5. Later, the clusters were integrated in a

single data file and used as input for Top Down Induction of Decision Trees (TDIDT) algorithm C4.5 [21] implemented in the data mining software Tanagra [28], using the cluster column as target attribute and the non-spatial attributes TMin, TMax, TAv as input attributes, with the non-normalized values of each of them, resulting in a decision tree that yields the rules that can be seen in Figure 6.

```
IF TAv < 9.85
AND TMin > -0.65
AND TMax < 13.5
THEN Cluster 0

IF TAv < 9.85
AND TMin < -0.65
THEN Cluster 1

IF TAv >= 9.85
AND TMin >= 7.2
THEN Cluster 2

IF TAv < 9.85
AND TMin > -0.65
AND TMax >= 13.5
THEN Cluster 3

IF TAv >= 9,85
AND TMin < 7,2
THEN Cluster 4
```

Figure 6. Description rules of the spatial clusters obtained using the proposed process.

The description rules allow differentiating clusters with 83.19% confidence. Cluster 0 has an average temperature lower than 9.85ºC, a minimum temperature higher than 0.65ºC below zero and a maximum temperature lower than 13.5ºC. This last value can distinguish cluster 0 from cluster 3, which has a maximum temperature higher than or equal to 13.5ºC.

On the other hand, cluster 1 can be characterized by its low average temperature and minimum temperature, with values lower than 9.85ºC and -0.65ºC respectively, while the same values are higher than 9.85ºC and 7.2ºC in cluster 2.

Finally, cluster 4 has average temperatures higher than 9.85°C and minimum temperatures lower than 7.2ºC.

In this case, due to the fact that all the attributes of the data have been used for the clustering and not only the spatial attribute, as in the previous proof of concept, the confidence was reduced to 83.19%, as can be seen in Figure 5, where clusters overlap spatially. However, each cluster is homogeneous enough to be described using only one rule for each of them.

V. CONCLUSION

A knowledge discovery process for description of spatially referenced clusters that works with any kind of clusters, i.e. regions, clusters of spatial objects and point patterns, and the description of the characteristic values of the attributes of its members using decision tree learning algorithms is proposed in this paper.

This knowledge discovery process provides a systematic way for business intelligence to obtain relevant information about automatically generated spatial clusters to be used in the decision making, in contrast to existing algorithms. Having a process that specifies the activities to conduct the analysis in a systematic way makes the clustering task more flexible for information mining engineers and decision-makers.

Two proofs of concept of the process using real data have been shown using two different kinds of spatial clusters: the first one uses an algorithm to generate regions, and the second one discovers clusters of spatial points.

In future works, the behavior of the process on data with many non-spatial attributes will be investigated, as well as the influence of choosing different spatial clustering algorithms.

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 Technological National University, 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.

REFERENCES

[1] García-Martínez, Ramón, Paola Britos, and Dario Rodríguez. "Information mining processes based on intelligent systems." International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer Berlin Heidelberg, 2013.

[2] Gopal, Ram, James R. Marsden, and Jan Vanthienen. "Information mining—Reflections on recent advancements and the road ahead in data, text, and media mining." (2011): 727-731.

[3] Mennis, Jeremy, and Diansheng Guo. "Spatial data mining and geographic knowledge discovery—An introduction." Computers, Environment and Urban Systems 33.6 (2009): 403-408.

[4] Kataria, Poonam, and Navpreet Rupal. "Mining Spatial Data & Enhancing Classification Using Bio-Inspired Approaches."International Journal of Science and Research (IJSR). 2012.

[5] Brimicombe, Allan J. "A dual approach to cluster discovery in point event data sets." Computers, environment and urban systems 31.1 (2007): 4-18.

[6] Guo, Diansheng. "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)." International Journal of Geographical Information Science 22.7 (2008): 801-823.

[7] Yang, Xiankun, and Weihong Cui. "A novel spatial clustering algorithm based on delaunay triangulation." International Conference on Earth Observation Data Processing and Analysis. International Society for Optics and Photonics, 2008.

[8] Zhong, Caiming, Duoqian Miao, and Ruizhi Wang. "A graph-theoretical clustering method based on two rounds of minimum spanning trees." Pattern Recognition 43.3 (2010): 752-766.

[9] Deng, Min, et al. "An adaptive spatial clustering algorithm based on Delaunay triangulation." Computers, Environment and Urban Systems 35.4 (2011): 320-332.

[10] Liu, Qiliang, et al. "A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity." Computers & Geosciences 46 (2012): 296-309.

[11] Shah, Glory H., C. K. Bhensdadia, and Amit P. Ganatra. "An empirical evaluation of density-based clustering techniques." International Journal

of Soft Computing and Engineering (IJSCE) ISSN 22312307 (2012): 216-223.

[12] Nisa, Karlina Khiyarin, Hari Agung Andrianto, and Rahmah Mardhiyyah. "Hotspot clustering using DBSCAN algorithm and Shiny web framework." Advanced Computer Science and Information Systems (ICACSIS), 2014 International Conference on. IEEE, 2014.

[13] Santoso, Aries, and Karlina Khiyarin Nisa. "Cloud Computing Application for Hotspot Clustering Using Recursive Density Based Clustering (RDBC)." IOP Conference Series: Earth and Environmental Science. Vol. 31. No. 1. IOP Publishing, 2016.

[14] Popat, Shraddha K., and M. Emmanuel. "Review and comparative study of clustering techniques." International Journal of Computer Science and Information Technologies 5.1 (2014): 805-12.

[15] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[16] Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.

[17] 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.

[18] Martins, Sebastian, Patricia Pesado, and P. García-Martínez. "Information Mining Projects Management Process." Proceedings 28th International Conference on Software Engineering & Knowledge Engineering. Pág. 2016.

[19] Silver, Bruce. "BPMN Method and Style, with BPMN Implementer's Guide: A structured approach for business process modeling and implementation using BPMN 2.0." Cody-Cassidy Press, Aptos, CA 450 (2011).

[20] Hinneburg, Alexander, and Daniel A. Keim. "An efficient approach to clustering in large multimedia databases with noise." KDD. Vol. 98. 1998.

[21] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[22] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[23] Ho, Tin Kam. "Random decision forests." Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. Vol. 1. IEEE, 1995.

[24] Ho, Tin Kam. "The random subspace method for constructing decision forests." IEEE transactions on pattern analysis and machine intelligence 20.8 (1998): 832-844.

[25] Delaunay, Boris. "Sur la sphere vide." Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7.793-800 (1934): 1-2.

[26] Complete US City Data – All In One Place. Sample Dataset. Online: https://www.uscitieslist.org/. Consulted on November 28, 2016.

[27] Aurenhammer, Franz. "Voronoi diagrams—a survey of a fundamental geometric data structure." ACM Computing Surveys (CSUR) 23.3 (1991): 345-405.

[28] Rakotomalala, Ricco. "TANAGRA: a free software for research and academic purposes." *Proceedings of EGC*. Vol. 2. 2005.

[29] National Institute of Agricultural Technology (INT A). Daily Data. SIGA – Agrometeorological Information and Management System. Argentina (2016). On-Line: http://siga2.inta.gov.ar/en/datosdiarios/. Consulted on June 6, 2016

[30] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11.1 (2009): 10-18.