Journal of Geomatics Vol. 13, No. 1, April 2019

A Spatial data mining approach applied in urban planning

Mohammed Midoun[1*] and Hafida Belbachir[2]
[1] Dept. Informatique Université des Sciences et de la Technologie Mohamed Boudiaf Oran USTO Oran, Algeria
[2] Dept. Informatique Université des Sciences et de la Technologie Mohamed Boudiaf Oran USTO
*Email: midoun.mohamed@univ-usto.dz

(Received: Apr 09, 2018; in final form: Apr 24, 2019)

Abstract: The nature of spatial information generates a set of problems of incompatibility with the principles of data mining. Spatial data mining is an extension of data mining that considers the interactions in space. It involves various techniques and methods in various areas of research. It takes into account the specificities of spatial information such as spatial relationships that can be topological, metric or directional. These relationships are implicit and difficult to represent. A Bayesian network is a graphical model that encodes causal probabilistic relationships among variables of interest, which has a powerful ability for representing and reasoning and provides an effective way to spatial data mining. Moreover, spatial data cubes allow storage and exploration of spatial data. They support spatial, non-spatial and mixed dimensions. A spatial dimension may contain vector and raster data. The spatial hierarchies can represent topological relationships between spatial objects. In this article we propose to use Bayesian networks for knowledge discovery in spatial data cubes. The goal of our approach is first to consider spatial relationships in the data mining process, and secondly to benefit from the strength of the data warehouses to apply spatial data mining on different aggregation levels according to the topological relations between spatial data.

Keywords: Spatial data mining, Bayesian networks, Spatial warehouse Spatial Analysis

1. Introduction

Spatial data mining is an extension of data mining that considers the interactions in space. It takes into account the specificities of spatial information, such as spatial relationships and spatial dependence. Many studies have been done, where association rules, clustering, classification methods or Bayesian networks are used. A Bayesian network is a graphical model that encodes causal probabilistic relationships among variables of interest, which has a powerful ability for representing and reasoning and provides an effective way for spatial data mining.

On the other hand, spatial data cubes are cubes that contain dimensions or facts which are spatially referenced and can be represented on maps (Bédard et al, 2001) They allow the storage and exploration of spatial data. They support spatial, non-spatial and mixed dimensions. A spatial dimension can contain vector data or raster data. The nature of spatial information generates a set of problems of incompatibility with the principles of data mining. First, the spatial data is linked, while the methods of datamining consider that the data are independent. On the other hand, the spatial relationships are implied and are seldom stored in databases.

The spatial relations are multiple, they may be topological (adjacency, intersection...) or metric (distance) and the analysis can be mono or multi thematic. This makes it difficult to choose the correct spatial relationship.

To represent the spatial relationships in relational databases, we can use the spatial joint index or contiguity matrix. Another approach is to model the spatial information into spatial data cubes. A spatial data cube is an ideal environment for data mining, it allows analysis and spatial queries on several levels of spatial aggregation. Several works on data mining on spatial data cubes were made. However, few studies have applied Bayesian networks on spatial data cubes. This, due to the complexity of spatial data sets.

Our major contribution is to propose a platform for the application of Bayesian networks on spatial data cubes for data mining purposes. To represent spatial relationships, we use a spatial hierarchy of vector layers that will respect the topological relationships between spatial objects. The spatial aggregation will be used to calculate the measures and then apply data mining on different levels of the spatial hierarchy.

The main interest of our contribution is to use Bayesian networks to apply spatial data mining on different levels of aggregation of spatial hierarchy. This by considering the spatial relationships.We will use spatial analysis to confirm the validation of our approach and view the results on a map.

In the next section we give an overview of some existing works pertaining to spatial data mining and spatial data cubes. Then we define our approach and we propose a framework of spatial data mining based on Bayesian networks. The results and evaluation of our approach will be discussed in experiments section. Finally, we end this paper with some conclusions.

2. Relevant literature

Spatial data mining is the application of data mining techniques to spatial data. It can be defined as the discovery of interesting, implicit and previously unknown knowledge from large spatial data bases (Bédard et al, 2001). The main objective of the spatial data mining is to discover relationship and characteristics that may exist implicitly in spatial databases. It has been used in various fields like remote sensing, medical imagery and visual data mining. Spatial Data Mining extends relational data mining with respect to special features of spatial data, like mutual influence of neighboring objects by certain factors (topology, distance, direction). Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

Many works have been proposed for spatial data mining, they relate to the various tasks of data mining, such as classification (Ester et al, 1997; Warrender and Augusteijn, 1999), association rules (Kamber et al, 1997) , or clustering (Han et al, 2001) .

Moreover, the application of Bayesian Networks for spatial data mining and knowledge discovery was introduced by (Han et al, 2001). Bayesian networks provide a coherent framework of representation and reasoning for spatial problems. The process of spatial data mining based on Bayesian networks includes two parts, one is structure learning, and the other is learning the parameters of the network. Many studies have focused on the learning of structure (Lam and Bacchus, 1994; Huang et al, 2004), and many others on the study of algorithms and methods of learning parameters (Oniśko et al, 2001; Feelders and Van der Gaag, 2006).

As for a spatial data mining method, Bayesian networks can be used for spatial knowledge representation, spatial classification, spatial clustering, and spatial prediction (Huang and Yuan, 2007). Several studies have been conducted: (Porwal et al, 2006) used Bayesian network classifiers for mineral potential mapping (Liebig, et al, 2009), developed an algorithm that can be applied to large trajectory collections (Walker et al, 2005), proposed Spatial Bayesian learning algorithms for geographic information retrieval, and (Li et al, 2012) proposed a bayesian method for assessing vulnerability to natural disasters to catastrophic risk.

Data warehouses are databases of information dedicated to the analysis and decision making (Kimball et al, 1996). A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process (Inmon, 1996). Spatial data warehouse is data warehouse where some dimension members or some facts are spatially referenced and can be represented on a map. Spatial data warehouses contain geographic data, for example, satellite images, and aerial in addition to non-spatial data.

A number of studies have been conducted for spatial data mining in spatial data cubes. They relate in particular to the use of association rules, classification methods, and exploitation of raster databases (Image)) (Han et al, 1998; Bédard et al, 2001).

The main difficulty in spatial data mining is the recognition of spatial relationships in databases. These spatial relationships are implicit and difficult to be represented. Several solutions have been proposed to solve this problem. The spatial relationships between objects in a spatial framework are often modeled by a contiguity matrix. A contiguity matrix can be representing a neighborhood relationship defined using the Euclidean distance or contiguity. Another solution proposed by Valduriez (Valduriez 1987) is to add a joint index to speed up the joints as part of a relational database. The extension to spatial data has been proposed by Zeitouni et al. in (Zeitouni et al, 2001). This extension consists of adding a third attribute that represents the spatial relationship between two objects More models of spatial relationships using hypergraphs are available in the literature.

Malinowski and Zimányi, (2005) propose to model topological relationships through spatial hierarchies of spatial data cube. They define the different types of spatial hierarchies. In addition, they classify topological relationships between hierarchical levels according to the procedures required for ensuring correct measure aggregation. A spatial data cube can include numerical measures and spatial measures and pointers to spatial objects at different levels of aggregation. Aggregation of spatial objects is not easy; it requires the use of a spatial hierarchy.

Few studies have applied Bayesian networks on spatial data cubes. Our contribution is to provide a methodology for the application of Bayesian networks on spatial data cubes. With the aggregation of spatial and non-spatial measures, our work allows to take in account the spatial relationships, including topological relationships between different objects, and perform a knowledge discovery in various aggregation levels of a spatial hierarchy.

3. Proposed approach

Figure 1 shows the approach we propose for the application of Bayesian networks on spatially referenced data. We use GIS data that we store in a database. These data are then integrated in a cube of spatial data, where the measurements are aggregated according to the different levels of the spatial hierarchy which corresponds to the topological relations of the spatial objects. Then we apply the Bayesian networks for data mining purposes to predict the progress of constructions in a housing program. The results obtained are then validated by comparing them with the results observed in the field

3.1 Dataset

In a Geographic Information System (GIS), there are vector data (geometric: point, line, polygon) and raster data (pixels). Our approach is based on GIS vector data. We get the data from multiple heterogeneous sources of spatial data (vector) and non-spatial data (attribute).

Then, we apply the pre-processing steps, such as, converting the vector data or adding spatial projection. We perform these pre-processing steps in a GIS environment, and then we build our spatial database.

.

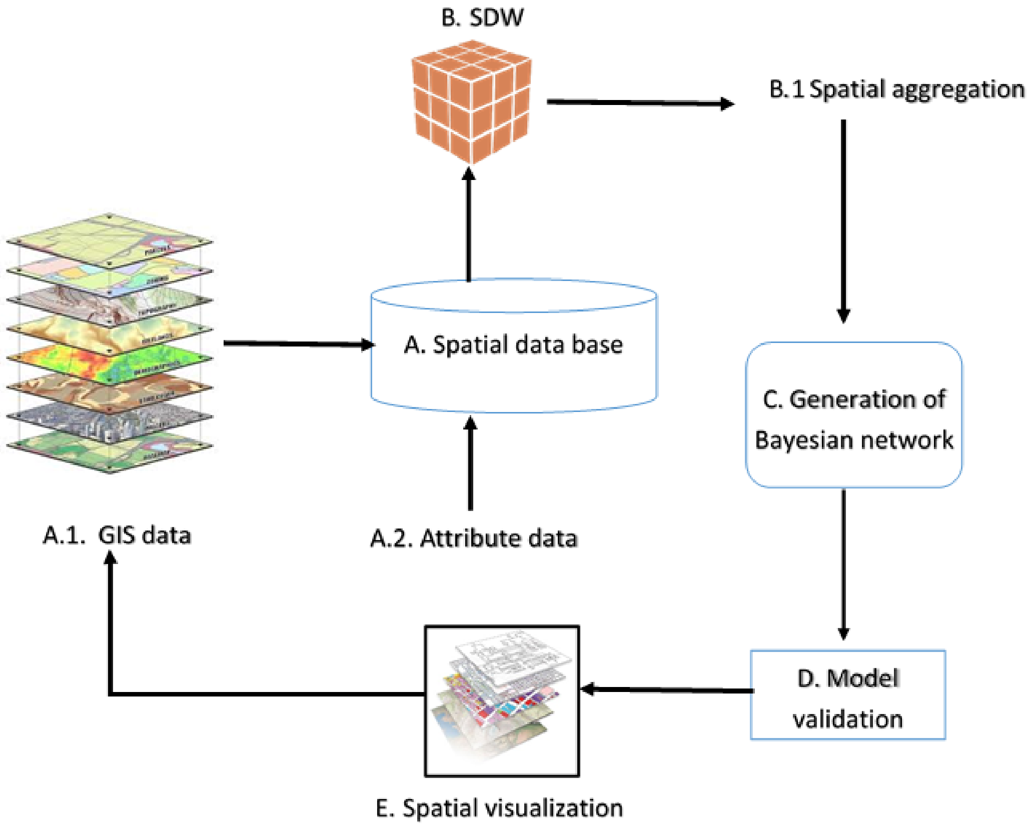

Figure 1: Proposed spatial data mining process

In the spatial data mining environment , we use not only GIS to manage and visualize spatial data, but also as a means of calculating spatial measures using spatial analysis techniques.

3.2 Spatial data cube

In spatial datawarehouse, spatial information can be integrated as dimensions or measures. Spatial data cubes are cubes for which members of dimensions or facts (via spatial measures) are spatially referenced and can be represented on maps (Bédard et al, 2001).

There are two types of spatial data cubes, vector cubes and raster cubes. They contain at least one dimension where some or all members are geometric.

In a data cube, data is organized in dimensions which describe in a natural way most of the attributes associated with the data of interest (Inmon 1992). The dimensions are in turn organized into hierarchies, with data aggregated at each level. As for the dimension hierarchies, topological relationships have hierarchical structures. these relations correspond to the hierarchical semantic relationships between spatial objects.

Therefore, our approach is based on the use of these topological relationships to add levels to the spatial hierarchy of our spatial data cube. The measures of the fact table will be aggregated and calculated according to each level of aggregation of the spatial dimension. They will represent the variables on which we apply Bayesian networks

3.3 Baysian networks

Bayesian networks are graphical models for defining probabilistic relationships between variables. An advantage of Bayesian networks is that they capture knowledge in a form people can understand intuitively, and which allows a clear visualization of the relationships involved.

Bayesian networks use a directed acyclic graph (DAG) to represent assertions of conditional relationships. The nodes in the graph represent the variables and the directed arcs define the conditional relationships.

The advantages of directed graphic models over undirected models are the notion of causality. Causality indicates that if an arc is directed from A to B in the network, then A causes B. Bayes' theorem is used to calculate causal inference about the variables. Bayes' theorem states:

$$\boldsymbol{P}(\boldsymbol{B} \mid \mathbf{A}) = \frac{\boldsymbol{P}(\boldsymbol{B} \mid \mathbf{A}).\ \boldsymbol{P}(\boldsymbol{A})}{\boldsymbol{P}(\boldsymbol{B})}$$

The construction of Bayesian networks is a difficult task, and the number of possible structures and parameters can be huge in this kind of structure.

Learning a Bayesian network from data involves two tasks: Estimating the probabilities for the conditional probability tables (learning parameters) and deriving the structure of the network.

The process of building the Bayesian network consists of three steps: variables definition, structure learning, and parameter estimation.

- Variable definition. Defines the relevant variables and the relationship between them.
- Structure learning. Determine the directions of all edges based on prior knowledge and the given data set. Structure learning of Bayesian networks is the key step to perform reasoning and predicting.
- Parameter estimation. It refers to define the conditional probabilities of the relationships. This step defines the conditional probabilities associated with each node.

As for a spatial data mining method, Bayesian networks can be used for spatial knowledge representation, spatial classification, spatial clustering, and spatial prediction.

Bayesian networks involves different search algorithms for constructing the network topology. The heuristic algorithms include K2, DAG, Hill Climbing (HC), and TAN (Tree Augmented Naive (Bayes)).

In our approach, we use Bayesian networks on measures of spatial data cube. The measures are calculated using aggregated levels of spatial hierarchy. Then, they will be discretized, and several Bayesian networks can be built from these measures.

3.4 Evaluation and validation

Once the Bayesian networks are built, each network must be evaluated. For this purpose, we compare the results obtained with real results observed in the field. The accuracy of the evaluation and the calculation of the Kappa index will allow us to evaluate the results obtained by our approach based on Bayesian networks.

3.5 Spatial analysis

Spatial analysis is a set of methods and tools which enable to understand, evaluate and interpret the spatial distribution of phenomena in order to discover and / or highlight the general rules of organization of space (Pumain and Saint-Julien, 1997). Spatial analysis can be applied to the interrogation of thematic, geometric and topological components of the spatial information contained in the GIS.

Once the Bayesian model is generated and validated, we integrate the parametric data of the Bayesian model into thematic GIS layers. This will not only allow spatial analysis to visualize the results on a map and compare them with the results obtained in the field, but will also validate the analysis of the generated Bayesian network.

4. Experiments

This section describes the experiment conducted to evaluate the proposed aproach. We apply Bayesian networks for Urban Planning in order to predict the progress of housing construction programs in Algeria.We use réal data stored in a spatial data cube.

4.1 Experimentation environnement

We applied our approaches under Windows environment, with SQL Server 2012 as database management system and ArcGIS 10.3 desktop as GIS.

4.2 Dataset

For the purposes of our experiments, we used a database comprised of vector GIS data prepared in ArcGIS environment, and non-spatial data from progress reports on housing construction programs in Algeria. Spatial data concerns the graphic representation of zones, islets, parcels and buildings. The non-spatial data relate to the progress rates of the constructions collected in the field as well as the information on each of the space objects.

These data were processed and integrated in spatial data cube modeled in snowflake schema as shown in figure 2. This cube has four dimensions: Report, Date, Phase, and Buildings, and one spatial hierarchy with four levels: Building / Parcels/ Islets / Area . This hierarchy represents the topological relationship between spatial objects classes.

There are 21 100 measures in the fact table of spatial data cube. They provide information on the progress of construction of buildings.

The measures are aggregated and calculated according to the spatial hierarchy, so as to apply Bayesian networks on different levels of aggregation of spatial data cube.

.

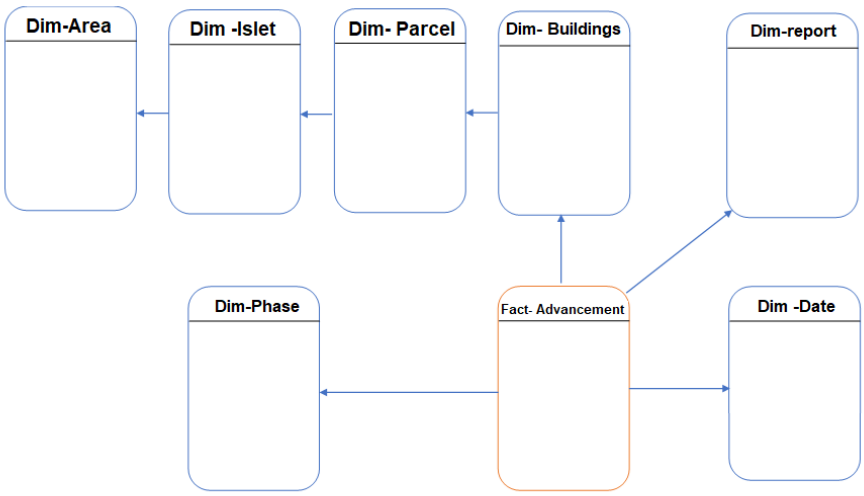

Figure 2: Spatial data warehouse

4.3 Bayesian networks

In our implementation we use the K2 algorithm to calculate the Bayesian network structure. K2 algorithm is the most famous score-based algorithm in Bayesian netowrk. It recovers the underlying distribution in the form of DAG efficiently. We begin by defining variables. The variables are described as following: Earthwork, Boundary marking, Concrete dosage, Verification of verticality, Verification of stability, Floor Coating, Partitionning, Coating, Window installation, Waterproofing, Painting.

The values of its variables are represented in the form of measures in the patial data cube. We perform the discretization and aggregation of these measures according to the levels of the spatial hierarchy.
After defining the domain variables and data preparation, we can obtain the structure of the Bayesian network and then we should compute the conditional probabilities of the relationship.

Figure 3 shows the structure of the Bayesian network applied for monitoring the building construction process. In this structure, we have 11 measurements, that are spread over six phases of construction (Infrastructure, Superstructure, Masonry, Coating, Water roofing and Painting). The progress of the construction phases is conditioned by the rate which represents each measure. The final rate of progress is conditioned by the rate of progress of the construction phases.

We apply Bayesian networks for each level of aggregation of spatial hierarchy (Buildings, parcels, islets). The Parametric results of the Bayesian network distribute the buildings, parcels and islets in six classes we have defined to represent the different stages of progress of the construction process.

4.4 Evaluation and validation

We selected 255 cases for testing the validity of the model. Table I shows a confusion matrix. It shows the results of the experiment. We compare the results obtained with real results observed in the field. The accuracy of the evaluation is 84.7% and the Kappa index is 0.848. The experimental results validate the proposed approach for spatial data mining. The parametric results obtained by the Bayesian networks represent the estimated rates of construction progress for the buildings, then aggregated for the parcels and islets according to the spatial hierarchy of the data cube. These estimated results are compared to the observed field results that represent the measures stored in the data cube.

Table 1: Confusion matrix generated by Bayesian network

		Observed results						
Evaluated results		***I***	***II***	***III***	***IV***	***V***	***VI***	***Total***
	I	***17***	*0*	*2*	*1*	*2*	*1*	*23*
	II	*2*	***20***	*3*	*1*	*0*	*1*	*27*
	III	*0*	*1*	***36***	*3*	*0*	*2*	*42*
	IV	*1*	*0*	*4*	***57***	*3*	*1*	*66*
	V	*1*	*0*	*2*	*1*	***30***	*3*	*37*
	VI	*0*	*2*	*0*	*1*	*1*	***56***	*60*
Total		21	23	47	64	36	64	255
Evaluation accuracy /%		**80,9**	**86,9**	**76,5**	**89,0**	**83,3**	**87,5**	

4.5 Spatial analysis

After validating the model, we integrate the parametric data of the Bayesian model in the GIS thematic layers: Buildings, parcels and islets . This allows us to perform spatial analysis, visualize the results on a map and compare them with the results observed in the field. The spatial analysis carried out were applied to the analysis of the progress of construction of buildings, plots and islets.

Figures 4 and 5 show the spatial distribution of buildings and islets. The construction process has six phases. In these figures, we visually compare the predictive results obtained with those observed in the field by the agents. This is done at several levels of aggregations that correspond to the topological relationships of spatial objects. Spatial analysis have enabled us to spatially visualize results and confirm the validation of our approach.

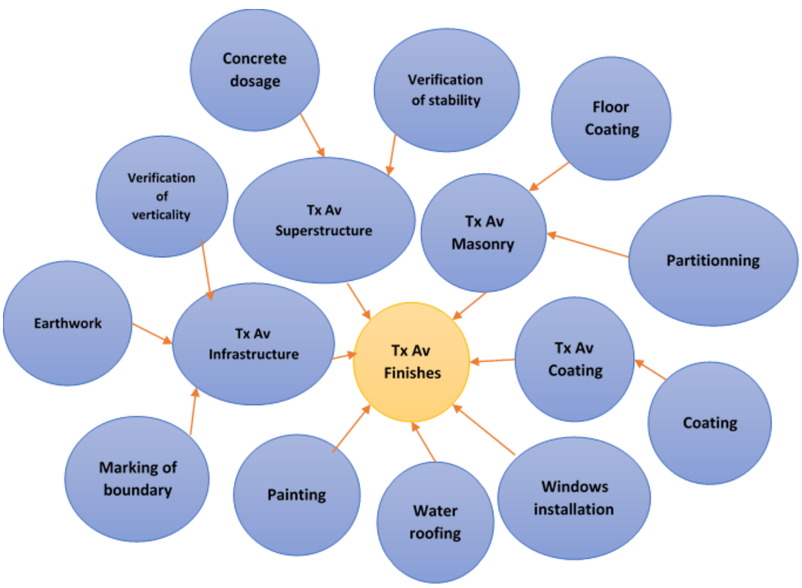

Figure 3: Bayesian network applied for urban planning

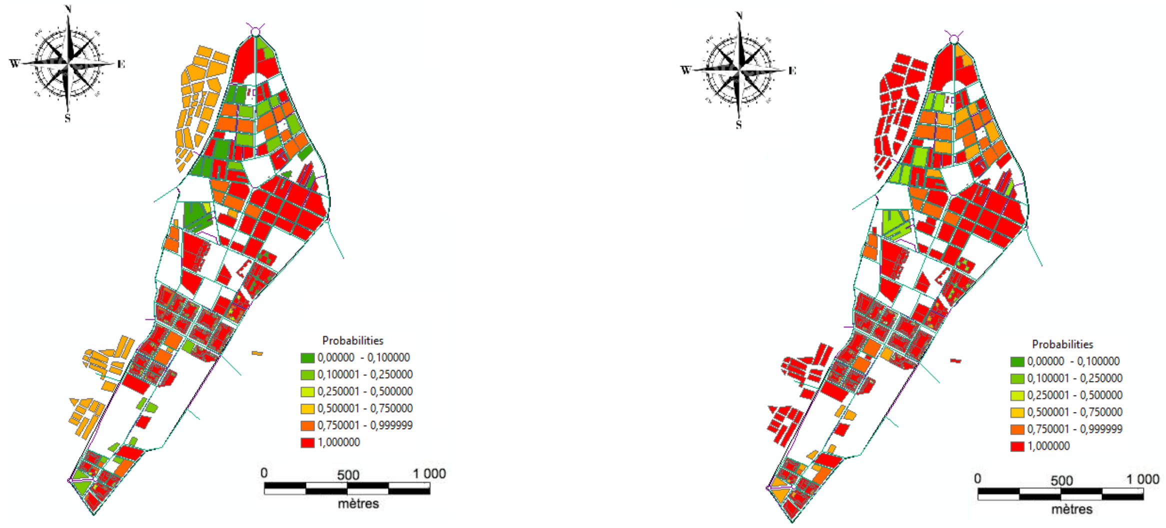

Figure 4: Spatial analysis of the distribution of buildings into classes according to the rate of progress of contruction and comparison of results with observed results

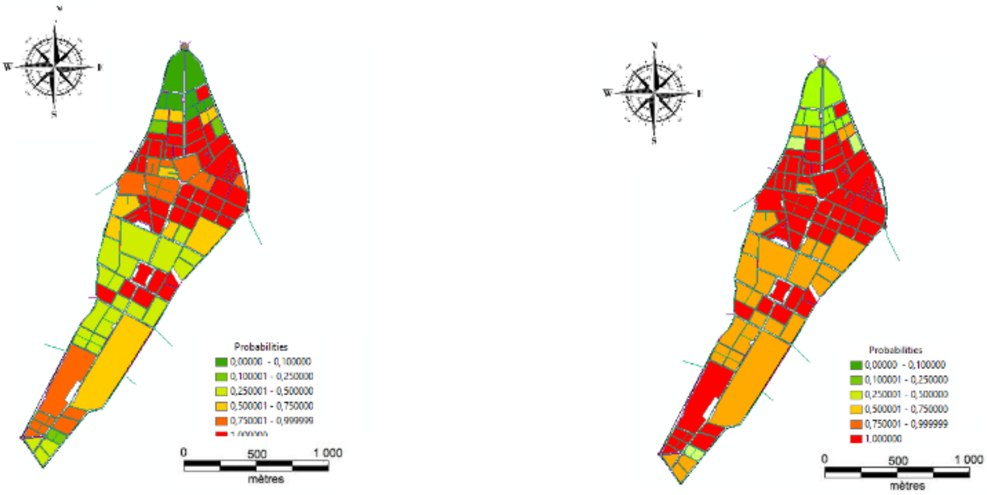

Figure 5: Spatial analysis of the distribution of islets into classes according to the rate of progress of contruction and comparison of results with observed results

4.6 Discussion

A s shown in Table I, the estimation accuracy was 84.7% and the Kappa index was 0.84 which is considered as a good result for prediction. On the other hand, using spatial analysis, the comparison of the results obtained with the results observed in the field validates our approach by a visual analysis. Data mining gives detailed results whereas spatial analysis gives a general description of the results. Spatial analysis is used to confirm visually the results obtained by data mining, but cannot be enough to give reliable results on its own.

We can conclude that the experimental results thus validate the feasibility of the proposed approach for knowledge discovery in spatial data. Moreover, the application of data mining on a spatial data cube allows a knowledge discovery about the different levels of aggregation of spatial hierarchy. Our approach allows not only to predict the construction progress of each building, but also the overall assessment of the construction process on the different islets and parcels of the study area.Another advantage of our method is to use GIS to visualize, validate and locate the results on a map. We can therefore say that our approach is a good way for spatial data mining in spatial data cubes.

5. Conclusion

Spatial data mining is an extension of data mining that takes into account the spatial relationships. Spatial relationships are difficult to be represented in databases. Few studies have used Bayesian networks for knowledge discovery in spatial data cubes. In this article, we first explain the concepts related to data mining and spatial data cubes. Then, we propose a framework for data mining in spatial data cubes using Bayesian networks. Furthermore, we showed a case study and used the experimental data to validate the applicability of Bayesian networks for spatial data mining. Consequently, we consider our approach as a good way to explore the spatial data.

The first interest of our approach is that it takes into consideration the spatial relationships including topological relationships. In addition, it allows knowledge discovery about the different levels of aggregation of spatial hierarchy. Another advantage of our method is to use spatial analysis and GIS to evaluate, visualize and locate the results on a map.

In conclusion, our study presents multiple perspectives, such as the development of a decision support tool that combines spatial analysis and Bayesian networks, or the development of new algorithms for Bayesian networks taking into account the spatial relationships in the process of knowledge discovery.

References

Bédard, Y., T. Merrett and J. Han (2001). Fundamentals of spatial data warehousing for geographic knowledge discovery. Geographic data mining and knowledge discovery, 2, pp. 53-73.

Ester, M., H.-P. Kriegel and J. Sander (1997). Spatial data mining: A database approach. International Symposium on Spatial Databases, Springer.

Feelders, A. and L. C. Van der Gaag (2006). Learning Bayesian network parameters under order constraints. International Journal of Approximate Reasoning, 42(1), 37-53.

Han, J., M. Kamber and A. Tung (2001). Spatial clustering methods in data mining: A survey. H. Miller and J. Hun, eds. Geographic Data Mining and Knowledge Discovery, Taylor & Francis, London, UK.

Han, J., N. Stefanovic and K. Koperski (1998). Selective materialization: An efficient method for spatial data cube construction. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer.

Huang, J., H. Pan and Y. Wan (2004). An algorithm for cooperative learning of Bayesian network structure from data. International Conference on Computer Supported Cooperative Work in Design, Springer.

Huang, J. and Y. Yuan (2007). Construction and application of Bayesian network model for spatial data mining. 2007 IEEE International Conference on Control and Automation, IEEE.

Inmon, W. (1992). Building the data warehouse, QED Technical Pub. Group.

Inmon, W. (1996). The data warehouse environment. In the Building the Data Warehouse, Wiley and Sons, New York.

Kamber, M., J. Han and J. Chiang (1997). Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes. KDD.

Kimball, R., M. Ross, W. Thornthwaite and B. Becker (1996). The Data Warehouse Toolkit., John Wiley&Sons. Inc., New York.

Lam, W. and F. Bacchus (1994). Learning Bayesian belief networks: An approach based on the MDL principle. Computational intelligence, 10(3), 269-293.

Li, L., J. Wang, H. Leung and S. Zhao (2012). A Bayesian method to mine spatial data sets to evaluate the vulnerability of human beings to catastrophic risk. Risk Anal, 32(6), 1072-1092.

Liebig, T., C. Körner and M. May (2009). Fast Visual Trajectory Analysis Using Spatial Bayesian Networks. ICDM Workshops.

Malinowski, E. and E. Zimányi (2005). Spatial hierarchies and topological relationships in the spatial MultiDimER model. British National Conference on Databases, Springer.

Oniśko, A., M. J. Druzdzel and H. Wasyluk (2001). Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. International Journal of Approximate Reasoning, 27(2), 165-182.

Porwal, A., E. J. M. Carranza and M. Hale (2006). Bayesian network classifiers for mineral potential mapping. Computers & Geosciences, 32(1), 1-16.

Pumain, D. and T. Saint-Julien (1997). L'Analyse spatiale. 1. Localisations dans l'espace. Paris: Armand Colin, coll. Cursus.

Valduriez, P. (1987). "Join indices." ACM Transactions on Database Systems (TODS), 12(2), 218-246.

Walker, A. R., B. Pham and M. Moody (2005). Spatial bayesian learning algorithms for geographic information retrieval. Proceedings of the 13th annual ACM international workshop on Geographic information systems, ACM.

Warrender, C. E. and M. F. Augusteijn (1999). Fusion of image classifications using Bayesian techniques with Markov random fields. International Journal of Remote Sensing, 20(10), 1987-2002.

Zeitouni, K. and L. S. Géographiques (2000). Fouille de données spatiales. Revue internationale de géomatique, (4/99).

Zeitouni, K., L. Yeh and M.-A. Aufaure (2001). Join indices as a tool for spatial data mining. Temporal, Spatial, and Spatio-Temporal Data Mining, Springer, pp.105-116.