USING SPATIAL DATA MINING TO DISCOVER THE HIDDEN RULES IN THE CRIME DATA

Karel, JANEČKA[1], Hana, HŮLOVÁ[1]

[1] Department of Mathematics, Faculty of Applied Sciences, University of West Bohemia

Univerzitni 22, 32300, Pilsen, Czech Republic
kjanecka@kma.zcu.cz

Abstract

The main of this research was to explore the possibilities of Oracle Spatial for spatial data mining. We used Oracle Spatial for finding of association rules in the crime data. In particular, we had data about robberies which happened in the Czech Republic in the year of 2008. We focused on robberies which were perpetrated by youth. The amount of crime data is increasing and needs modern and effective processing. The crime data contain both spatial and non-spatial part. It makes sense explore the crime data if there are some regional patterns. In our research the thematic crime data were offered by the Czech Police Headquarters. We obtained this source thematic data in many xls files. Extraction, transformation and loading of data into the Oracle database were the initial steps of the spatial data mining process. In addition, we applied the built-in functionality of Oracle Spatial for materialization of loaded data. In particular the spatial binning method was used. The last step was done in Oracle Data Miner. For finding of association rules the Apriori algorithm was applied. We received many important association rules. These rules show that the situation about the crime perpetrated by youth differs from region to region. The results of this research were offered to the Czech Police Headquarters. Consequently appropriate measures can be applied to remedy the situation in particular regions.

Keywords: CRISP-DM methodology, ETL process, Spatial Binning, Regional Patterns, Oracle Spatial

1. INTRODUCTION

Today we can hear and read more and more that young people perpetrate crime. It is a very dangerous fact. It is necessary to explore the reasons why it is so. Police is responsible for solving it. Police is collecting data which are next processed and analyzed, mostly in geographical information systems. The amount of this data is rapidly increasing. Due to this reason the using of data mining seems to be an adequate method for exploring of huge amount of crime data. But the thematic crime data are relative to positions on the Earth and therefore it is not possible to use only the methods of classical data mining. It is about using of both data and spatial data mining methods.

There are currently several methodologies for data mining which we can be used in many application fields. As an example we can mention the CRISP-DM methodology. This is an industry and tool neutral data mining process model and consists of six steps [1]:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

The relationships between particular phases are illustrated on figure 1. Spatial data mining is broadly used in geographical information systems, geomarketing, Earth observation, navigation and many other areas [2]. It is used for better understanding of relationships in data, for discovering of hidden relationships between spatial and attribute data and also for optimizing of spatial queries.

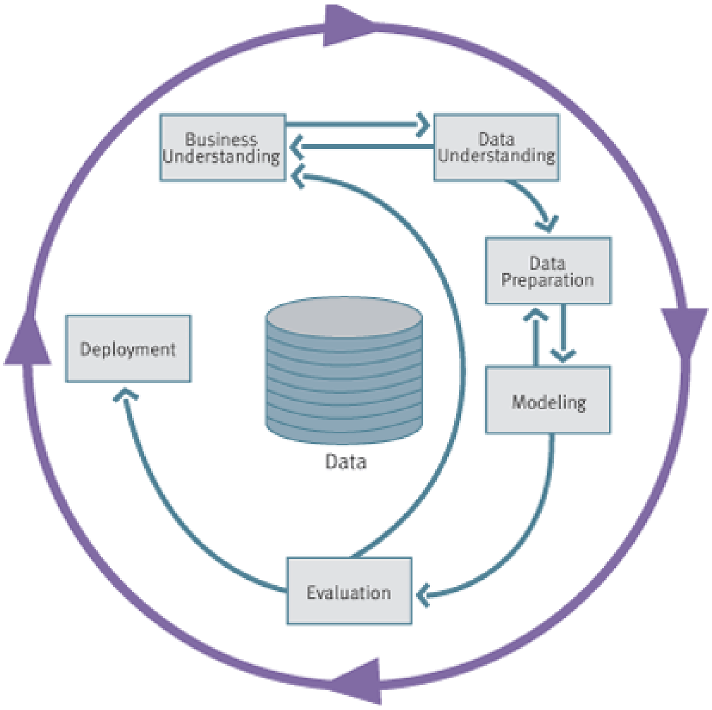


Fig. 1 The current process model for data mining provides an overview of the life cycle of a data mining project. It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks. [1]

Our research was inspired by previous researches done in the field of spatial data mining and crime data in which generally the principles of the CRISP-DM model are used [3, 4, 5]. In our research we also followed the CRISP-DM model principles. However, the referenced works didn't use only the spatial database management system for spatial data mining process. Therefore one of the main aims of this research was to explore the possibilities of Oracle Spatial for spatial data mining process. We believe that describing of hidden associations in the crime data in the form of association rules is very readable for crime specialists to be able make important decisions. Therefore we strongly focused on possibilities of Oracle Spatial for generating of association rules.

2. SPATIAL DATA MINING

Spatial data mining can be defined as follows:

"Spatial Data Mining (SDM) is a well identified domain of data mining. It can be defined as the discovery of interesting, implicit and previously unknown knowledge from large spatial data bases." [6]

The spatial data mining is more complicated than classical data mining due to the complexity of spatial data types, spatial relationships and spatial correlation among features. The spatial correlation means that the object of interest is influenced also by the neighbouring features. That is the reason why we have to also consider the attributes of "neighbours" during the spatial mining process. An effectiveness of many algorithms depends on the effective processing of relationships with surrounding.

2.1. Unsupervised Data Mining

An unsupervised data mining (UDM) is sometimes also called as "teaching without a teacher". It means that known proved values are not available. The both methods clustering and finding of association rules belong to this group of data mining methods. Due to the reason that known patterns are not available before application of these methods we can use them for description of relationships and found patterns in data. In our research we especially concentrated on the method of finding of association rules in combination with the spatial binning for detection of regional patterns in crime data.

2.2. Spatial Data Mining in Oracle Spatial

Oracle is a relational database management system with the advanced possibilities of data processing. Oracle Spatial supports also spatial analysis and mining in Oracle Data Mining (ODM) applications. ODM allows automatic discovery of knowledge from a database. Its techniques include discovering hidden associations between different data attributes, classification of data based on some samples, and clustering to identify intrinsic patterns. Spatial data can be materialized for inclusion in data mining applications. The spatial analysis and mining features in Oracle Spatial let as exploit spatial correlation by using the location attributes of data items in several ways: for binning (discretizing) data into regions (such as categorizing data into northern, southern, eastern, and western regions), for materializing the influence of neighbourhood, and for identifying collocated data items.

The original data, which included spatial and nonspatial data, is processed to produce materialized data. Spatial data in the original data is processed by spatial mining functions to produce materialized data. The processing includes such operations as spatial binning, proximity, and collocation materialization. The ODM engine processes materialized data (spatial and nonspatial) to generate mining results. [2]

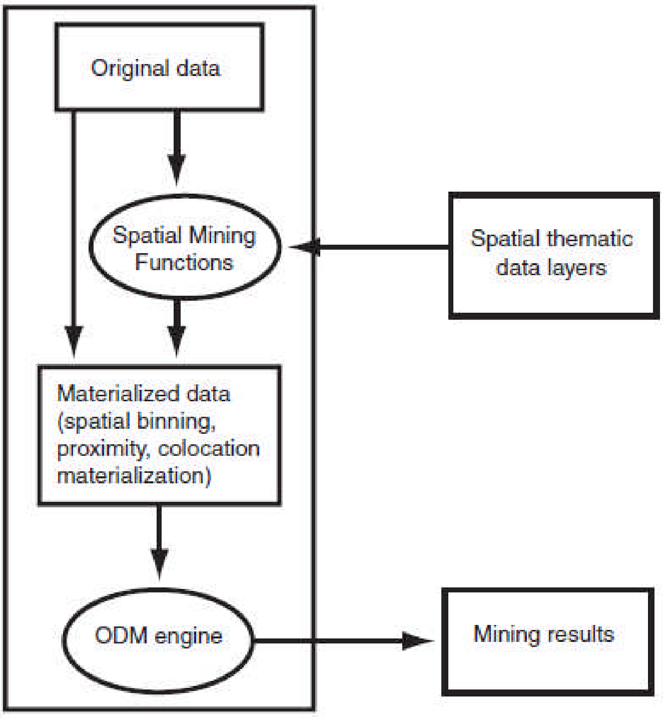

Original data
Spatial Mining Functions
Spatial thematic data layers
Materialized data (spatial binning, proximity, colocation materialization)
ODM engine
Mining results

Fig. 2 A scheme of the Spatial Mining process in Oracle Spatial. The spatial materialization could be performed as a preprocessing step before the application of data mining techniques, or it could be performed as an intermediate step in spatial mining. [2]

2.3. Spatial binning for detection of regional patterns

Spatial binning (spatial discretization) discretizes the location values into a small number of groups associated with geographical areas. The assignment of a location to a group can be done by any of the following methods:

- Reverse geocoding the longitude/latitude coordinates to obtain an address that specifies (for United States locations) the ZIP code, city, state, and country.
- Checking a spatial bin table to determine which bin this specific location belongs in [2].

We were applying ODM techniques to the discretized locations to identify interesting regional patterns and association rules in the crime data.

2.4. Association

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. Association rules are often used to analyze sales transactions but association modelling has also important applications in other domains as well [7].

3. INPUT CRIME DATA

The thematic crime data were offered by the Czech Police Headquarters – the department of prevention. Overall we got nearly 90 xls files including among other things the source data about robberies which happened in the year 2008. Table 1 gives information about the count of robberies in the year 2008 in the Czech Republic.

Table 1 Count of all facts and robberies which were registered in the Czech Republic in the year 2008

Count of district police departments	Count of registered facts	Count of robberies
196	223 036	3 435

There are 14 main administrative regions in the Czech Republic and many district police departments belong to each region as stated on figure 3.

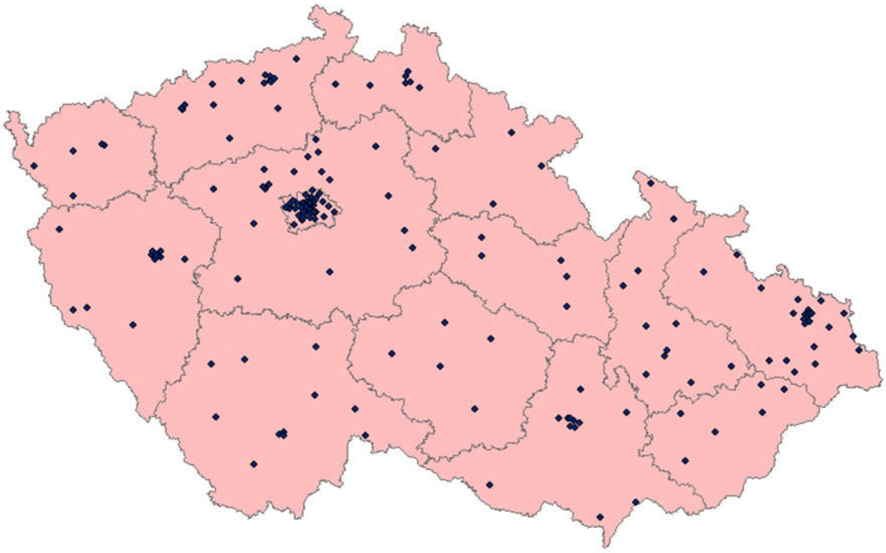


Fig. 3 The Czech Republic consists of the 14 main administrative regions. Dots represent locations of district police departments

The aim of the research was to try finding out some hidden associations in this crime data in particular main administrative regions. In particular we concentrated on robberies which were perpetrated by youth (up to 15 years of age; next as “youth<15”) and youth (between 15 and 18 years of age, according by the Czech law; next as “youth>15”) because this is a very negative social phenomenon.

4. PROCESSING OF INPUT DATA

To be able make spatial data mining we had to process all input data through Extraction – Transformation – Loading (ETL) process. ETL is a process that involves the following tasks:

- Extracting data from source operational or archive systems which are the primary source of data for the data warehouse;
- Transforming the data - which may involve cleaning, filtering, validating and applying business rules;

- Loading the data into a data warehouse or any other database or application that houses data [8].

4.1. Extraction

In addition to the aforementioned crime data, we used the locations of police departments of the Czech Republic organized in shp file. The data were extracted from the files of Land Identification Register – Basic Area Units (Czech Statistical Office [9]). In addition, boundary of the Czech Republic and boundaries of 14 main administrative regions were used.

4.2. Transformation

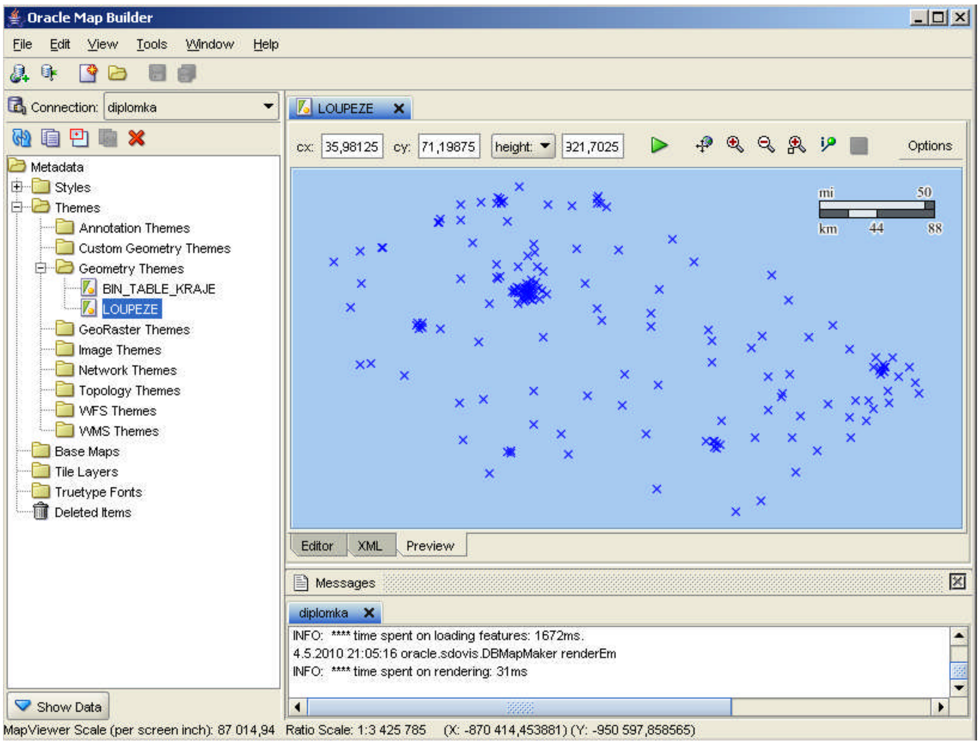
It should be stated that not all attributes were considered during the spatial data mining. From the source thematic data we filtered out only the following attributes for each robbery: a day (Monday, Tuesday …), time of day (1 – 6 a.m., 7 – 12 a.m., noon – 6 p.m., 6 – 12 p.m.), kind of thief in accordance with his age (as described in chapter 3) and alcohol (if the robbery was done after the influence of alcohol).

The above mentioned attributes dealing with robberies were added to the dbf table of shapefile with positions of police departments. Now, we had thematic information relative to some spatial position.

4.3. Loading

First of all it was necessary to load the extracted data into the Oracle database. The tables for storing of extracted and transformed data were created. The spatial attributes like boundaries of regions were modeled as objects and stored in SDO_GEOMETRY [2] columns. This data type offers to model spatial features in object-relational way. All geometry description of spatial feature is then stored in one cell in spatial table. There were created two main tables:

- `BIN_TABLE_KRAJE` (for storing of 14 regions; one `SDO_GEOMETRY` column for storing of regions' boundaries),
- `LOUPEZE` (for storing of data in shapefile with thematic information; one `SDO_GEOMETRY` column for storing of positions of police departments). Figure 4 illustrates the positions of district police departments.



Connection: diplomka

cx: 35,98125 cy: 71,19875 height: 321,7025

```
INFO:  **** time spent on loading features: 1672ms.
4.5.2010 21:05:16 oracle.sdovis.DBMapMaker renderEm
INFO:  **** time spent on rendering: 31ms
```

MapViewer Scale (per screen inch): 87 014,94 Ratio Scale: 1:3 425 785 (X: -870 414,453881) (Y: -950 597,858565)

Fig. 4 Visualization of district police departments' positions

The loading of data was done by using of the Oracle Map Builder application [18]. This application allows user to create, to maintain and to view spatial data and metadata. It allows also loading the shapefile into the Oracle database.

4.4. Application of Spatial Binning on the loaded data

As we needed to join the `BIN_TABLE_KRAJE` and `LOUPEZE` tables by using of the Spatial Binning method, we had to add some columns to the previous tables. In particular, the column `BIN` was added to the `BIN_TABLE_KRAJE` table and column `ID_BIN` to the `LOUPEZE` table.

The `SDO_SAM.BIN_GEOMETRY` function, documented in [2], performs operation related to spatial binning. In particular, it computes the most-intersecting tile for geometry. After applying of this function on the `BIN_TABLE_KRAJE` table fourteen bins were created and their identifiers were stored in the `BIN` column. Finally, for spatial assigning of each police department (stored in the `LOUPEZE` table) to the appropriate bin the `SDO_SAM.BIN_LAYER` [2] procedure was used.

Now, we had materialized data (see figure 2) ready for spatial data mining.

5. FINDING OF ASSOCIATION RULES

The Apriori algorithm was used for finding of association rules. This algorithm calculates the probability of an item being present in a frequent itemset, given that another item or items is present. The Apriori algorithm calculates rules that express probabilistic relationships between items in frequent itemsets For example, a rule derived from frequent itemsets containing A, B, and C might state that **IF** A and B are included in a transaction, **THEN** C is likely to also be included. An association rule states that an item or group of items implies the presence of another item with some probability. Unlike decision tree rules, which predict a target, association rules simply express correlation [7].

5.1. Antecedent and Consequent

The **IF** component of an association rule is known as the *antecedent*. The **THEN** component is known as the *consequent*. The antecedent and the consequent are disjoint; they have no items in common. Oracle Data Mining supports association rules that have one or more items in the antecedent and a single item in the consequent [7].

5.2. Metrics for association rules

Two main metrics are used to influence the build of an association model - *support* and *confidence*. Support and confidence are also the primary metrics for evaluating the quality of the rules generated by the model. Additionally, Oracle Data Mining supports *lift* for association rules. These statistical measures can be used to rank the rules and hence the usefulness of the predictions [7]. In our work we used only support and confidence.

Support

The support of a rule indicates how frequently the items in the rule occur together. Support is the ratio of transactions that include all the items in the antecedent and consequent to the number of total transactions.

Confidence

The confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction. Confidence is the conditional probability of the consequent given the antecedent. Confidence is the ratio of the rule support to the number of transactions that include the antecedent [7].

Example of computing support and confidence metrics for the rule R_1 **IF** A and B **THEN** C for the following transactions:

Transaction ID	Items
1	(A, B, C)
2	(D, A, B)
3	(A, C, D)
4	(A, B, D)

Probability of antecedent (A, B) and consequent (C) is 25% because only one transaction contains items A, B and C. Therefore the support for the rule R_1 is 25%. Transactions 1, 2 and 4 contain items A and B. It means that the confidence for the rule R_1 is 33%.

5.3. Preparation of transactional data

Unlike other data mining functions, association is transaction-based. Due to this fact it was necessary to transform our data stored in table `LOUPEZE`. To be possible to find out some regional patterns we created fourteen tables including transactional data. Each table contained transactional data for one particular administrative region. Each "transactional" table contained four columns:

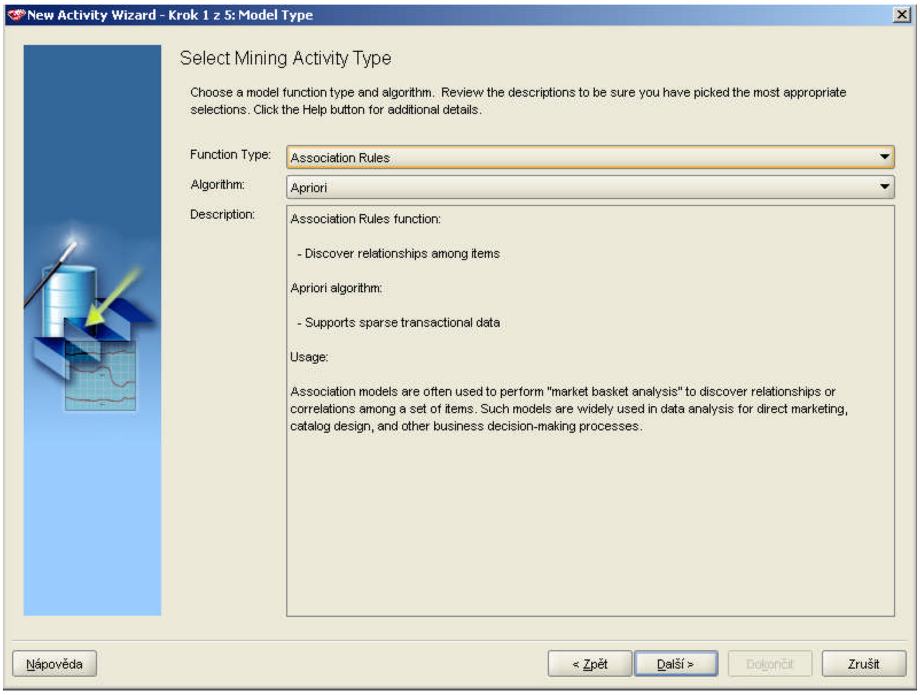
- `id` – an identifier of transaction
- `id_police_department` – an identifier of district police department,
- `police_department` – a name of district police department,
- `code_of_act` – possible values are defined in chapter 4.2.

Values in the `id` column were generated by database management system. Each triplet of values for the remaining columns was transformed from the `LOUPEZE` table.

5.4. Generating of association rules

The Oracle Data Miner [10] (ODM) application was used for retrieving of hidden relationships and association rules in the crime data in the form **IF** A AND B **THEN** C. There is a five-step wizard we can use for finding of association rules [11] in ODM. In the first step, we must select Function Type and used Algorithm. We used Association rules and the Apriori algorithm as theoretically described above. The choice of Function type and Algorithm is illustrated on Figure 5.

Next step was the selecting of source table with transactional data. All fourteen tables with the transactional data about the robberies were subsequently chosen. An identifier of each transaction was also set up in this step. In particular, the attribute `code_of_act` was used. In the third step of this process next attributes from "transactional tables" were chosen. Next, a name of the new table for storing of found association rules was entered. Finally, the association rules were generated and stored in the new table. It was necessary to enter some appropriate values for Support and Confidence parameters. We tried to set up these values repeatedly and explore the retrieved association rules. Finally, to get some predicative result we set up the Support parameter on value 30% for each region. The Confidence parameter was set up on value 20% for each region. In last step of wizard association rules were retrieved.



New Activity Wizard - Krok 1 z 5: Model Type

Select Mining Activity Type

Choose a model function type and algorithm. Review the descriptions to be sure you have picked the most appropriate selections. Click the Help button for additional details.

Function Type: Association Rules

Algorithm: Apriori

Description: Association Rules function:

- Discover relationships among items

Apriori algorithm:

- Supports sparse transactional data

Usage:

Association models are often used to perform "market basket analysis" to discover relationships or correlations among a set of items. Such models are widely used in data analysis for direct marketing, catalog design, and other business decision-making processes.

Fig. 5 The Apriori algorithm was used for retrieving of association rules

Table 2 Association rules for particular regions with the highest value of Support and Confidence for the robberies perpetrated by "youth<15"

Region	Association Rule	Confidence	Support

South Bohemia	**IF** robbery on Monday AND in time 1 p.m. – 6 p.m.
	THEN committed by "youth<15"
		100%	45%

Pilsen	**IF** robbery perpetrated by youth AND in time 7 p.m. – 12 p.m. **THEN** committed by "youth<15"	80%	33%

Karlovy Vary	**IF** robbery **on Tuesday AND in time 1 p.m. – 6 p**.m.
	THEN committed by "youth<15"
		33%	20%

Usti nad Labem	**IF** robbery on Monday AND in time 1 a.m. – 6 a.m.
	THEN committed by "youth<15"
		81%	50%

Liberec	**IF** robbery on Thursday AND in time 7 a.m. – 12 a.m.
	THEN committed by "youth<15"
		50%	11%

Hradec Kralove	**IF** robbery on Wednesday AND in time 1 a.m. – 6 a.m.
	THEN committed by "youth<15"
		100%	25%

Central Bohemia	**IF** robbery perpetrated by youth AND in time 7 a.m. – 12 a.m **THEN** committed by "youth<15"	50%	13%
Prague (the capital)	**IF** robbery perpetrated by youth AND on Tuesday **THEN** committed by "youth<15"	47%	18%
Vysocina	**IF** robbery perpetrated by youth AND on Sunday **THEN** committed by "youth<15"	100%	33%

Olomouc	**IF** robbery on Monday AND in time 7 p.m. – 12 p.m.
	THEN committed by youth
		100%	45%

Moravian-Silesian	**IF** robbery on Thursday AND in time 7 a.m. – 12 a.m.
	THEN committed by "youth<15"
		100%	62%

South Moravian	**IF** robbery on Wednesday AND in time 7 a.m. – 12 a.m.
	THEN committed by "youth<15"
		100%	61%

5.5. Regional patterns

It was recognized that there are differences in the retrieved association rules in different administrative regions. We concentrated on the children's and young's criminality (robberies). For example, in the South Bohemian Region, if the robbery happened on Monday in time 1 p.m. – 6 p.m., then it was committed by child in 45%! The Support for this association rule was 100%, Confidence 45%. Against this fact, the situation in the neighboring Pilsen region was quite different. The table 2 contains selected examples of association rules for particular regions with the inclusion of the values of both Support and Confidence parameters.

More retrieved association rules and their description can be found in [12].

6. CONCLUSION

One of the aims of this research was to explore the possibilities of Oracle database including Spatial for spatial data mining. This topic is very wide. We tried to apply the method of finding association rules in spatially materialized data. For materialization of spatial and thematic (crime) data the spatial binning was used. The results of this work were taken over by the Czech Police Headquarters. Some information rising from the research was really surprising. For example, the children's criminality in Prague, the capitol, is not so much high as it is generally supposed to be. The work also presented the possible way how to process the huge amount of data which are collected by police. The retrieved association rules can help to adapt appropriate measures to remedy the situation in particular region. From the technical point of view we tried to describe all process of spatial data mining with the crime data. We described the ETL-processing of crime data and spatial data and the transactional structure of corresponding tables which are a base for retrieving of association rules.

REFERENCES

[1] Cross Industry Standard Process for Data Mining. http://www.crisp-dm.org/index.htm, Cited: 10/2010

[2] Murray, Ch. Oracle Spatial Developer's Guide 11g Release 1. Oracle, 2009.

http://download.oracle.com/docs/cd/B28359_01/appdev.111/b28400.pdf, Cited: 03/2010

[3] McCue, C. (2007) Data Mining and Predictive Analysis: Intelligence gathering and crime analysis. Butterworth-Heinemann.

[4] Cherukuri, K.; Muralikrishna, Reddy, V. Spatial and Collateral Data Mining for Crime Detection and analysis.

http://www.gisdevelopment.net/application/miscellaneous/me05_183.htm, Cited: 11/2009

[5] Nath, S., V. Crime Pattern Detection Using Data Mining. Oracle

http://www.oracle.com/technetwork/database/options/odm/overview/crime-patterns-snath-odm-134695.pdf, Cited: 11/2009

[6] Witte, E. Spatial Data Mining.

http://www1.in.tum.de/teaching/ws01/CBP-Hauptseminar/Presentations/SpatialDataMining-Pres.pdf, Cited: 03/2010

[7] Oracle Data Mining Concepts 11g Release 1. Oracle, 2008.

http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129.pdf, Cited: 04/2010

[8] ETL-Tools.Info. http://etl-tools.info/en/bi/etl_process.htm, Cited: 10/2010

[9] Uzemne identifikacni registr UIR-ZJS, Cesky statisticky urad.

http://www.liberec.czso.cz/csu/rso.nsf/i/prohlizec_uir_zsj, Cited: 03/2010

[10] Oracle Data Miner. http://www.oracle.com/technetwork/database/options/odm/downloads/index.html, Cited: 03/2010

[11] Haberstroh, R. Oracle Data Mining Tutorial. Oracle, 2008

http://www2.tech.purdue.edu/cit/Courses/CIT499d/ODMr%2011g%20Tutorial%20for%20OTN.pdf, Cited: 04/2010

[12] Hulova, H. Aplikace vybranych metod prostoroveho dolovani dat v databazovych systemech. Diploma Thesis. Pilsen, 2010