APPLICATION OF DATA MINING ON PUBLIC ENVIRONMENTAL DATA

Vejmelka M., Jalovecký J., Urubková A.
Západomoravská vysoká škola Třebíč o.p.s.;
miloslav.vejmelka@cez.cz, jalovecky@via-alta.cz, aurubkova@zmvs.cz

ABSTRACT:
This article aims to summarize the legislative and technical framework for waste management, and on this basis to analyze the fulfillment of these issues in individual communities of the Vysočina region. The region has implemented integrated waste management system (ISNO). Waste management is regulated by legislation in Act No. 185/2001 Coll.: On waste and amending some other Acts. Based on the legislative and professional requirements mentioned issues will be established criteria for the collection and association rules for data evaluation. Expected primary source of data are web pages of Vysočina region municipalities. This data will be supplemented by other relevant data from other sources (Czech Statistical Office, Republic surveyor, web resources). The obtained statistical data set will then be evaluated based on the methods of data mining and specifically according to the methodology GUHA - procedure 4 ft. The data will be processed using the software Ferda Data Miner. The resulting hypotheses are interpreted using GIS.

KEYWORDS: Vysočina Region, Waste Management, Data Mining, GUHA, association rules, geodata, hypothesis.

1. Legislative and professional framework

The general source of environmental law is the law č.17/1992 Coll. as other laws. The purpose of this Act is to define the basic concepts of the principles of environmental protection and deals with the legal entities and individuals to protect and improve the environment.
Other laws that address the legal regime of the environment are: Act. No. 254/2001Sb. the Water Act. No. 86/2002 Coll. Air Protection Act. No. 289/1995 Coll. on Forests Act. č.114/1992 Coll. on nature and landscape.
The issue of waste is regulated by Act No. 185/2001 Coll. on waste and amending some other Acts, as amended.

2. Analysis of hypotheses

2.1 Short description of the problem

Subjects of examination are all villages, which lies in the Region. It is an examination of the objects that meet the condition in Equation 1

$$OBCE\_KV_{(KODNUTS=CZ063)} \subset OBCE \qquad \text{[Equation 1]}$$

2.2 Data bases

Data bases

Data source name	Location of stored data	Software for preprocessing	Note
KV_GEODATA	ArcČR 500 – digital geographic database 1:500 000, ARCDATA PRAHA, s.r.o. [1]	ArCGIS, MS Excel	Shapefile: Counties, Districts, Municipalities, Roads
KV_UIR	Territorial Identification Register basic settlement units, http://www.czso.cz [2]	MS Excel, MS Access	Data for municipalities in Vysočina region
KV_SD	List of waste disposal. Taken from: http://www.asekol.cz/sberna-mista/sberne-dvory.html?region=1113 [3]	MS Excel, MS Access	Data for municipalities in Vysočina region
KV_ODPADY	Competitions results. Taken from: http://www.tridime-vysocina.cz/vysledky-souteze/ctvrtleti/ [4]	MS Excel, MS Access	Data from competition database for municipalities in Vysočina region
KV_ELODPAD	Electronic waste collection points. Taken from: https://system.ekolamp.cz/index2.php ?stranka=systemsberu/seznam_sberny ch_mist_vypis [4]	MS Excel, MS Access	Data for municipalities in Vysočina region
KV_OBCE	Municipalities data. Taken from: http://mesta.obce.cz/ [5]	MS Excel, MS Access	Data from the mandatory information about the village

Tab. 1: List of Data bases

2.3.1 Preprocessing of data bases, determination of association rules

Determine the distance from each municipality to the seat of waste disposal for individual districts of *Vysočina region.*

All operations for the detection of these data were obtained using the potential of ArcGIS.

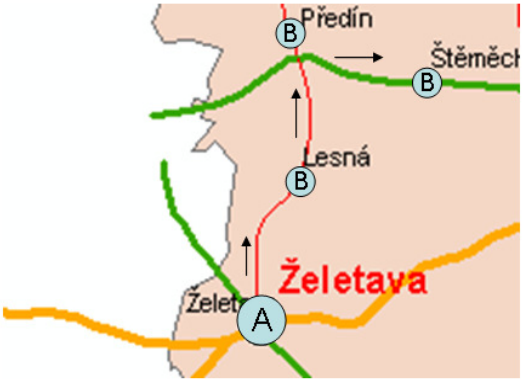

Figure 1: Principle of function Point Distance (A = OUTPUT-FID, B = INPUT_FID)

Results of Point Distance procedure were sorted by KODOB (town code) generated for each distance SD_OKRES - OBCE_OKR (by code number collection point) and minimum, maximum and arithmetic mean of minimum was calculated for each municipality.

Integration of data tables into one table.
Selected tool for Data Mining (Ferda Data Miner) needs to have the input data table in first normal distribution. To get the data from individual data tables in one their merger was carried out by key field (KODOB). Duplicate data were deleted, missing data were supplemented with data Boolean logic value of zero. Merging data into one table (KV_DATA) was done in MS Access tables were merged contents KV_GEODATA, KV_UIR, KV_SD, KV_ODPADY, KV_ELODPAD, and KV_OBCE SD_KV. **

$$** \; SD\_KV = \sum_{n=1}^{OkresyKV} KV\_OKRES \qquad \text{[Equation 1]}$$

3. Selection of GUHA tools

3.1 GUHA

GUHA procedure is a program that manages the individual procedures. Input of GUHA procedure is analyzed data and simple entry of set of potentially interesting relationships. It can be association rules or extensively more general relations.
Output are all simple relationships, it is such a link, which is true in the data analyzed and does an easy way of relations included in the output.

The most widespread procedure GUHA procedure ASSOC (GUHA procedure 4-ft miner) [6, 7], which seeks relations generalizing association rules. ASSOC entry procedure is the data matrix.
Rows of the matrix correspond to the observed objects and columns of the matrix correspond to attributes which describe the objects. Each attribute can take finitely many values.

ASSOC procedure works with the association rules following form:

$$Ant \approx Suc \qquad \text{[Equation 2]}$$

where Ant and Suc are boolean attributes that are derived from the columns of the analyzed matrix.

Boolean attribute is called the antecedent Ant, Suc is a Boolean attribute succedent.

Ant and Suc attributes are derived from the column matrix. The basic Boolean attribute used in the association rule is called literal, negated the basic Boolean attribute is a negative literal.

Derived attributes are Boolean conjunctions or disjunctions of literals.

3.2 Specification of GUHA procedures

3.2.1 Set of formulas
The following formulas were specified from input data set KV_DATA. These formulas will be further explored by GUHA procedure and selected according to the chosen quantifier.

Antecedent				Succedent
		VYMERA	≈	EFSB

				OBAKT	≈	EFSB
		MIN	&	MAX	≈	EFSB
KANAL	&	VODA	&	PLYN	≈	EFSB
PAPIR	&	PLAST	&	SKLO	≈	EFSB

Tab. 2: The set of studied formulas

3.2.2 Quantifier

Quantifier is a function defined over four-field table against which the hypothesis is verified. For our example is most often used four ft. quantifier, *better-informed implication*. This quantifier has one input parameter (*p*). With is is for tested whether at least (p) percent of the object fulfilling antecedent also fulfills succedent.

3.3 Ferda Data Miner

3.3.1 Product description

Ferda Data Miner was founded as a software project on Mathematics and Physics at Charles University in Prague under the lead of Prof. Mgr. John Rauch, MD. (University of Economics in Prague). The project was supposed to clarify and simplify the procedure for use in Data mining. The original procedure consisted of a gradual lowering of different applications and parameter setting, which did not at first sight connection. The program is modular in design so that it can be easily expanded to other boxes. We managed to create an environment that streamlines and simplifies the user work with the system and offers more possibilities for research and dissemination. [8]

3.3.2 Process of generating hypotheses

The procedure for generating hypotheses is described only investigated for the first formula. For other formulas similar procedure was elected with only slight differences.

Studied formula

The following formula was verified:

$$VYMERA \approx EFSB$$

Expressed in words: there is a relationship between the value and effectiveness of the collection area of the village. The aim is to find all the hypotheses that confirm (refute) the validity of this formula.

Assembly of the examined chain:

First have to be prepared chain (chain), which leads from the data source to final GUHA procedures.

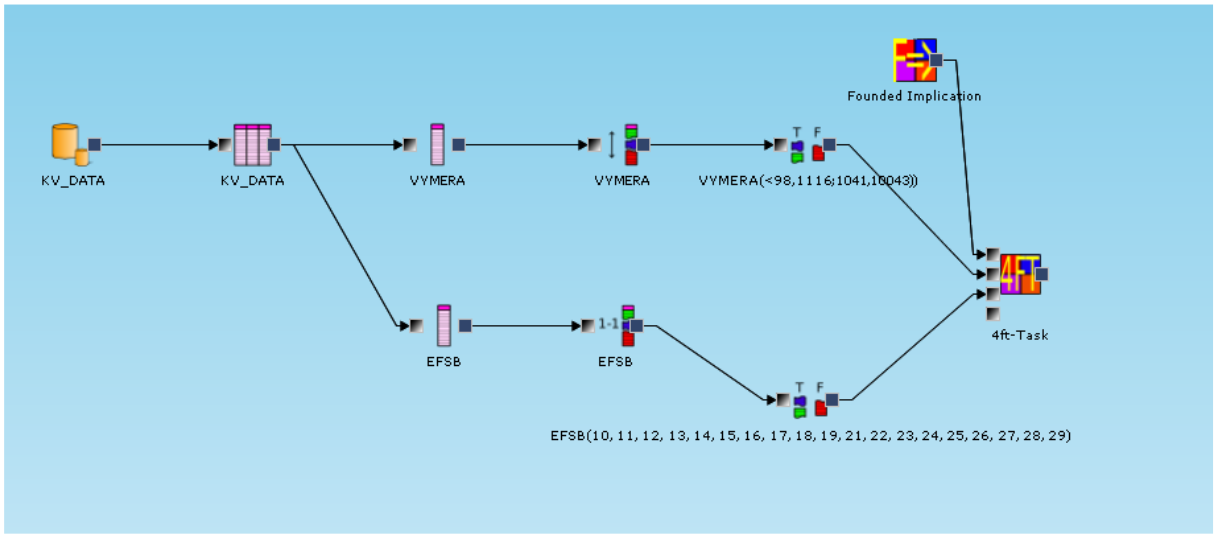

Figure 2: Ferda Data Miner – assembly of chains

Frequency distribution was calculated for columns VYMERA and EFSB by using the possibilities of Data Miner.

Design and optimization of hypotheses.
In this section, hypotheses were generated for items EFSB value that represents the most significant frequency (values 0 and 50 were not considered).
The strongest hypothesis was as follows:
Relevant four-field table is here:

KV_DATA	*EFSB*	*¬EFSB*
VYMERA	307	208
¬VYMERA	107	82

The hypothesis of parameters can be determined following conclusion: With the probability of 60% on collection efficiency (10 to 29) represents the major part villages with an area from 98.1 to 1041. Thus, each municipality that fits into that area, has a decent chance that the collection efficiency under current conditions will be between 10 to 29 percent.

3.3.3 Other hypotheses

Formula

$$OBAKT \approx EFSB$$

Result:

KV_DATA	*EFSB*	*¬EFSB*
OBAKT	407	281
¬OBAKT	7	9

Conclusion: Municipality with a population of between 22 to 5,142 (22 - 220) shows 59% probability that the collection efficiency will be in the range of 10 to 29%.

Formula

$$MIN\ \&\ MAX \approx EFSB$$

Result:

KV_DATA	*EFSB*	*¬EFSB*

MIN	171	114
¬*MIN*	269	150

KV_DATA	*EFSB*	¬*EFSB*
MAX	135	102
¬*MAX*	305	162

Conclusion: The village with the distance to the collection point (MIN 3054 - 9162 and 21378 to 30541) and (MAX 23 730 -32 458) has a probability of 60% or 57% efficiency that her collection will be in the range of 10 to 29%.

Formula

$$PLYN \,\&\, \text{VODA} \,\&\, KANAL \approx EFSB$$

Result:

KV_DATA	*EFSB*	¬*EFSB*
PLYN	440	150
¬*PLYN*	0	0

KV_DATA	*EFSB*	¬*EFSB*
VODA	365	208
¬*VODA*	75	56

KV_DATA	*EFSB*	¬*EFSB*
KANAL	290	184
¬*KANAL*	150	80

Conclusion: The villages that have established public water supply and do not have sanitation show a probability of 64% or 61% that their collection efficiency will be in the range of 10 to 29%. The existence of gas distribution in the village has no effect on this hypothesis.

Formula.

$$PAPIR \,\&\, \text{P}LAST \,\&\, SKLO \approx EFSB$$

Result:

KV_DATA	*EFSB*	¬*EFSB*
PAPIR	228	216
¬*PAPIR*	102	48

KV_DATA	*EFSB*	¬*EFSB*
PLAST	375	164
¬*PLAST*	65	100

KV_DATA	*EFSB*	¬*EFSB*
SKLO	349	127
¬*SKLO*	91	137

Conclusion: Towns which collect plastic and glass but do not collect paper showing probability 73%, 70% and 61% that their collection efficiency will be in the range of 10 to 29%.

4. Conclusion

For the final assessment of the situation was chosen GIS environment, specifically ArcGIS tools. Into attribute tables were added (Join) values, on which the hypotheses were tested. All conclusions drawn from the hypothesis had set filter over all municipalities Vysočina region.

Result in a graphic expression of the filter is shown in Figure No. 3.

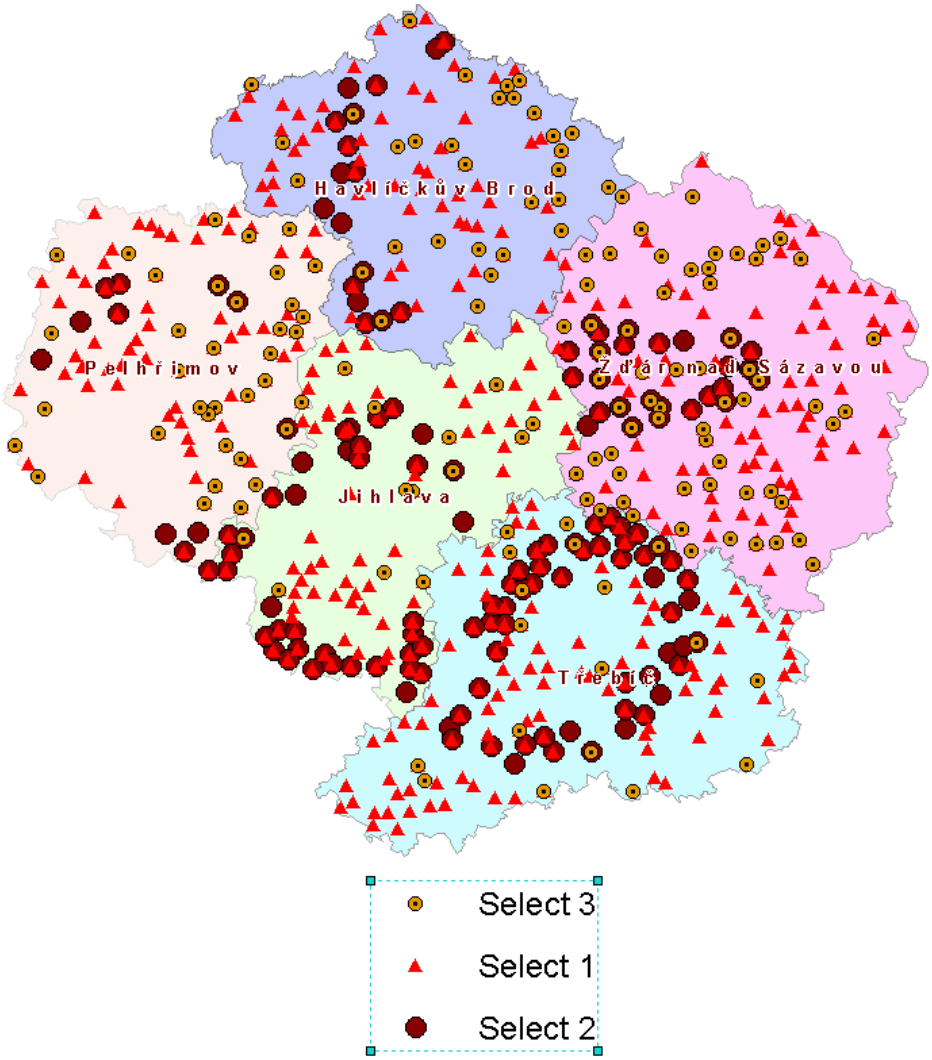

Figure 3: Graphical representation of hypotheses in ArcGIS

Overlapping symbols indicate that the municipality complied with more choices at a time. On the graphical display is easy to see the villages, which may require further analysis and adjustment of waste management.

LITERATURE:

[1] ČESKO. ARCDATA PRAHA, *Geografická data*, [online], [cit. 2012-02-10]. Dostupné z: http://www.arcdata.cz/

[2] ČESKO, ČESKÝ STATISTICKÝ ÚŘAD, [online], [cit. 2012-02-10]. Dostupné z: http://www.czso.cz/csu/redakce.nsf/i/home

[3] ČESKO, ENVI WEB, *Odpady*, [online], [cit. 2012-02-10]. Dostupné z: http://www.enviweb.cz/odkazy/odpady

[4] ČESKO, KRAJ VYSOČINA, *Čtvrtletní výsledky HLAVNÍ části krajské soutěže obcí „My třídíme nejlépe“ 2011*, [online], [cit. 2012-02-10]. Dostupné z: http://www.tridime-vysocina.cz/vysledky-souteze/ctvrtleti/

[5] ČESKO, VEŘEJNÁ SPRÁVA ON-LINE, *Obce On-line*, [online], [cit. 2012-02-10]. Dostupné z: http://mesta.obce.cz/

[6] HÁJEK, P., HAVRÁNEK, T., CHYTIL, M. K.: *Metoda GUHA*. Praha, Academia, 1983, 314 s.

[7] HÁJEK P, HAVEL I, T, CHYTIL M. K.: *Metoda GUHA automatického vyhledávání hypotéz*. Kybernetika 2, 1966, s. 31-41

[8] Projekt LISp Miner, [online], [cit. 2012-02-10]. Dostupné z: http://lispminer.vse.cz/.