УДК 004.6

A. Berko, I.Hlaholeva,
Lviv Polytechnic National University,
Information Systems and Networks department

APPLICATION OF DATA MINING AVERAGES TO FORECAST THE USE OF LAND RESOURCES

This article describes the procedures of data mining based on prediction of time series for land cadastre data. Principles, required for the development of the method of forecasting using time serious are examined. Mathematical model for serious prediction is developed. The task of prediction of land resources using in Striy District Lviv Region is technically realized.

Keywords: land cadastre, data mining, forecasting.

У статті описано процедури інтелектуального аналізу даних на основі прогнозування часових рядів для даних земельного кадастру. Розглянуті положення, необхідні для розробки методу прогнозування з використанням часових рядів. Розроблено математичну модель для прогнозування рядів, а також технічно реалізовано задачу прогнозування використання земельних ресурсів у Стрийському районі Львівської області.

Ключові слова: земельний кадастр, інтелектуальний аналіз даних, прогнозування.

Introduction

Efficiency of the land use is an important factor in the processes of decision-making related to land management. The land cadastre data contain a good many of records which form essential potential for efficiency enhancement of the land resources use. One of the ways to implement such potential is to forecast possible options of the resource use – a process of forming probable indicators of a defined object, which is directed towards discovery and examination of possible alternatives its future development. Forecasting is an important link between theory and practice in many branches of social activity, especially in management of sustainable land use. Forecasting of land use processes makes it possible to solve the tasks of efficient and sustainable land exploitation, and ensures land supply and demand balance.

For the last decades, a large number of spatial data were collected, which are now stored at the Central Board of the State Committee on Land Resources. The volumes of those data are so large that their analysis takes substantial time, financial, and technical costs, although, the necessity of such analysis is in fact obvious, since ‘raw data’ like these contain knowledge which may be used during approval of decisions to determine further processes of land use during a long time lapse.

One of today’s most promising averages to solve tasks on the efficient use and processing of large volumes of data are technologies based on methods of data mining.

Data Mining, DM, represents a process of discovery of hidden or not explicitly presented records but suitable for use large data sets. Data mining involves various mathematical methods for discovery of objective laws, ties and trends which exist in such data. Such objective laws may generally not be discovered during traditional data scanning, as their ties are too complicated, or may not be defined due to excess data volumes [1]. Data mining (DM) is a multidiscipline area which originated and has developed based on sciences like, applied statistics, pattern recognition, artificial intelligence, database theory, and other similar tooling. Data mining is a part of a process of knowledge discovery in data bases (KDD). It enables to unlock the essence of hidden data dependencies, elicit mutual impacts between object properties, information about which is stored in databases, extract objective laws peculiar for a specific data set [1]. Basic data mining methods and algorithms include, but are not limited to: neural networks, decision trees, symbol rules, nearest neighbour methods and k-nearest neighbour methods, support vector machines, Bayes networks, linear regression, correlation-regression analysis, hierarchical methods of cluster analysis, non-hierarchical methods of cluster analysis, including algorithms of k-average and k-

median, methods of search for association rules, that is to include Apriori algorithm; method of limited exhaustion, evolution programming ad genetic algorithms, various methods of data visualization [1].

Most analytical methods used in data mining are well-known mathematic algorithms and methods. New side of their application is a feasibility of using these averages when solving one or another specific problem of data processing, such feasibility based on appearance of new capacities of hardware and software averages. It also should be mentioned that most data mining methods were developed as a part of an artificial intelligence theory [2].

Analysis of recent publications and researches

Time indices are the key factor to deal with when solving forecasting tasks. In the middle 90-s of the last century, a radically new and rather strong class of algorithms was created to forecast time series. A great part of work on methodology research and models' testing was conducted by two statisticians, G.E.P. Box and G.M. Jenkins [3]. Since then, construction of similar models and receipt of forecasts on their basis have been called the Box-Jenkins method. This family involves few algorithms most well-known and used of them is the ARIMA (autoregressive integrated moving average). It is embedded practically in every specialized package for forecasting. The classic ARIMA has no independent variables. Models are supported only by information stored in the history of forecasted series. Modern scholarly literature often mentions options of the ARIMA models, which make it possible to include independent variables. Unlike the aforementioned methods of temporal series forecasting, the ARIMA methods provide for no concrete model to forecast the given temporal series. Only a general class of models is given, which describe a time series and which make it possible to somehow reflect current variable value through its previous values. After that, the algorithm by setting-up its internal parameters, implements a self-selecting of most suitable forecasting model. Matching and ties between the Box-Jenkins models [2], [3] form a hierarchy of models which may be logically described as a following sequence

$$AR(p)+MA(q)\rightarrow ARMA(p,q)\rightarrow ARMA(p,q)(P,Q)\rightarrow ARIMA(p,q,r)(P,Q,R)\rightarrow ...$$

Each element of this sequence is defined by a model of a specific type, however specific correlations are held between these models. According to [3], such models are describes as follows

Autoregressive order model $p-AR(p)$**.** The model is as follows:

$$Y(t)=f_0+f_1\cdot Y(t-1)+f_2\cdot Y(t-2)+...+f_p\cdot Y(t-p)+E(t), \quad (1)$$

where $Y(t)$ – a dependent variable at t time moment;

$f_0, f_1, f_2, ..., f_p$ – estimated parameters;

$E(t)$ - - error caused by impact of variables not considered in the given model [3].

Key task when applying this model is to set values of estimated parameters $f_0, f_1, f_2, ..., f_p$. There are currently in use few ways of solving such task, the most common of which is a search for values of such parameters through a system of the Yule-Walker equations [2]. To build this system, a calculation of autocorrelation function values is required. More simple way of obtaining parameters $f_0, f_1, f_2, ..., f_p$ is to calculate them by a least square method.

MA(q) **– model with a moving average** of *q*. The model is as follows:

$$Y(t)=m+e(t)-w_1\cdot e(t-1)-w_2\cdot e(t-2)-...-w_p\cdot e(t-p), \quad (2)$$

where $Y(t)$-a dependent variable at *t* time moment;

$w_0, w_1, w_2, ..., w_p$ - estimated parameters.

Model with a autoregressive moving average ARMA(p,q). ARMA(p,q) [2] averages a model of p autoregressive compounds with q of moving averages. More detailed presentation ARMA(p,q) model involves *AR(p)* and *MA(q)* models*:*

$$X_t = c + e_t + \sum_{i=1}^{q} \theta_i \cdot e_{t-i} + \sum_{i=1}^{p} \phi_i \cdot X_{t-i}, \quad (3)$$

Generally, value of e_t error is considered to be independent equally distributed random values obtained by normal distribution with a zero average: $e_t \sim N(0, \sigma^2)$, where σ^2 — dispersion. The assumption may be relaxed, but this may result in a change of model properties. For instance, model behaviour varies considerable if independence and equal error distribution are not assumed.

ARIMA (p,d,q) model. A task on analyzing of a temporal series with a complicated structure often involve models of *ARIMA(p,d,q)* [3] (Autoregressive Integrated Moving Average) of *(p,d,q)*, order which simulate different situations which occur during analysis of stationary and non-stationary series. Depending on the analyzed series, *ARIMA (p,d,q)* model be transformed up to the *AR(p)*, *MA(q)* moving average model or *ARMA (p,q)* mixed mode. During transition from non-stationary series to stationary series value of d parameter which defines a difference order is taken equal to 0 or 1, which means this parameter has only integer values. Generally, there is a limited choice between *d = 0* and *d = 1*. However, the research workers lose a situation when d parameter may take fractional values.

ARFIMA(p,d,q) model. For examination of fractional values difference order, foreign scholars, that is to say, C. W. Granger, J. R. Hosking, P. M. Robinson, R. Beran, suggested in their works a new class of models *ARFIMA(p,d,q)* [3] (F: fractional), which leaves room for a non-integral parameter d and autoregressive fractional integrated process of moving average. Such series has its specific features: self-similarity, fractal, slowly descending correlation. Forecasting of temporal series with the help of *ARFIMA(p,d,q)* model is more promising for forecast accuracy enhancement.

ARIMA (p,d,q)(P,D,Q)S model. This model is described by an expression *ARIMA (p,d,q)(P,D,Q)S* [3], where: *p* - autoregressive summands; *d* – differences; *q* – moving average summands; *P* – seasonal autoregressive summands; *D* – seasonal differences on S interval; *Q* – seasonal moving average summands.

Tasks and objectives of the article

Basic objective of the article is to develop methods of land use forecasting with the use of data mining. The tasks defined by this objective are as follows:

- research and support of feasibilities for data mining methods and means usage in the processes of forecast estimate formation for land use,
- simulation of processes to analyze cadastre data with the use of time series,
- development of an order to form forecast estimates for land use basing on models based on time series application,
- testing of decisions developed during the work and based on data of land cadastre of Stryi district, Lviv region.

Basic results of the researches

Application of Microsoft Analysis Services. Data mining means is a standard tooling for solution of analysis tasks in database management system of Microsoft SQL Server. When applied, they create additional opportunities for the use of data accumulated during a significant period of time. Microsoft DBMS analysis technologies are based on application of the aforementioned defined and described models. Such model represents a set of metadata reflecting dependencies, rules and objective laws supported by raw data. However, the model structure defines a set of key attributes, and its content is formed by statistical and data generic values. This approach provides for no possible changes, or construction of new models. This is the way how data mining training means process is implemented.

One of the models using DBMS Microsoft SQL Server data mining is a model for the analysis of time series. Microsoft time series algorithm implements algorithms of regression optimized to forecast continuous values in time. Unlike the other algorithms, like decision trees, the time series model requires no additional new data columns to forecast a trend. With the help of time series model trends may be

forecasted based only on reference data set used for model creation. When forecasting, new data may be entered to the model and automatically added to the trend analysis procedures.

As a part of DBMS Microsoft SQL Server 2005 time series algorithm (Microsoft) uses a single ARTXP algorithm. It was optimized for short-term forecasting, that's why it forecasted a subsequent probable value in the series. Starting with the SQL Server 2008 version, the time series algorithm (Microsoft) together with the ARTXP algorithm uses another algorithm, ARIMA. Algorithm ARIMA was optimized for long-term forecasting.

Microsoft time series use a default algorithm combination for the analysis of objectives laws and preparation of forecasts. The algorithm studies two separate models of the same data: one model uses the ARTXP algorithm, and the other uses the ARIMA algorithm. Later, algorithm unites the results of both models to form the best forecast for variable number of temporal slices. The ARTXP algorithm better suits for short-term forecasting, hence, when beginning a forecast series one should rely more on this algorithm. But as temporal slices used for the forecasting move ahead to the future, the ARIMA algorithm becomes more useful.

One may manage a combination of algorithms, bringing the emphasis in temporal series to short-term or long-term forecasting. Starting with the SQL Server 2008 Standard, one of the following modes may be recommended for temporal series algorithm (Microsoft):

- use of only the ARTXP algorithm for short-term forecasting;
- use of only the ARIMA algorithm for long-term forecasting;
- use of combination of two algorithms (default).

When using combined model, the temporal series algorithm (Microsoft) unites the two algorithms as follows:

- to form two first forecasts only the ARTXP algorithm is always used;
- after the first two forecasts, a combination of ARIMA and ARTXP algorithms is used;
- with further forecasting steps, share of ARIMA algorithm at forecasting increases until complete abandonment of the ARTXP use;
- to manage the moment of algorithm combination, speed of the ARTXP algorithm share reduction and increase of the ARIMA algorithm share, one may implement by changing parameters PREDICTION_ SMOOTHING.

Both algorithms may show the impacts of seasonal factors to data on various levels. For instance, in the middle of annual cycles, monthly cycles of data change may occur. In order to determine seasonal cycles, one may enter data on periodicity, or set a mode of periodicity automatic discovery.

When preparing data to be used in the study of any data mining model one must understand requirements of a specific model and methods of data use.

Every forecasting model must contain a set of options, that is, a table column indicating time slices, or other series where changes occur. Other models' options may be a text field or any other identifier, e.g. consumer code or transactions code. However, temporal series model must always contain date, time or other unique numeric value as a set of options.

Requirements for temporal series model:

- A separate key time column. Each model must contain one numeric column or "date" column to be used as a set of options to determine temporal slices used by the model. "Key time" column data type may be datetime or any other numeric type. However, this column must contain continuous values which must be unique foe every series. A set of options for temporal series model may not be stored in two columns, like "Year" column or "Month" column.
- A forecasted column. Each model must contain at least one forecasted column which will be a base for the algorithm to build temporal series model. Data type for the forecasted column must be with continuous values. For instance, it may be forecasted the way with the lapse of time change numeric attributes, like income, sales volumes or temperature. However, a column containing discrete values, like consumer status or educational level may not be used as a forecasted column.

- An optional key series column. Each model has an additional key column containing unique values to identify the series. This optional key series column must contain unique values. For instance, one model may contain data about sales of many product models, provided there is only one record for each product title in each temporal slice.

After the model study, the results are stored as a set of objective laws which may be investigated or be a base for forecasting[4].

Forecasting of lands with Microsoft Analysis Services. Information about the use of lands of Stryi district, Lviv region is stored at the Board of the State Committee on Land Resources, form "6-zem"which is a report on available lands and their allocation according to their owners, land users, holdings and types of economic activity for the year.

When forecasting land use with the help of time series algorithm, the information for the last five years was used, and with SQL Server Management Studio a data mining model was built.

In the process of simulation, a set of restrictions as regards composition and methods of data supply subject to data mining were considered. In particular, time series model must contain one numeric column or "date" column. "Year" field was chosen as such column. Each model must contain at least one forecasted column which will be a base for the algorithm to build time series model. The forecasted column must have a data type with continuous values. In order to build a forecast the following table field were chosen:

- total area of agricultural lands;
- area of agricultural holdings (total);
- lands in the course of ameliorative construction and soil fertility restoring;
- polluted agricultural lands (total);
- forests and other forested areas (total);
- built-up lands (total);
- water;
- lands intended for nature protection purposes;
- public lands;
- recreation lands;
- industry lands;
- lands intended for historical and cultural purposes.

The process of creation of data mining model with the use of time series algorithm was implemented with the help of the following DMX-request:

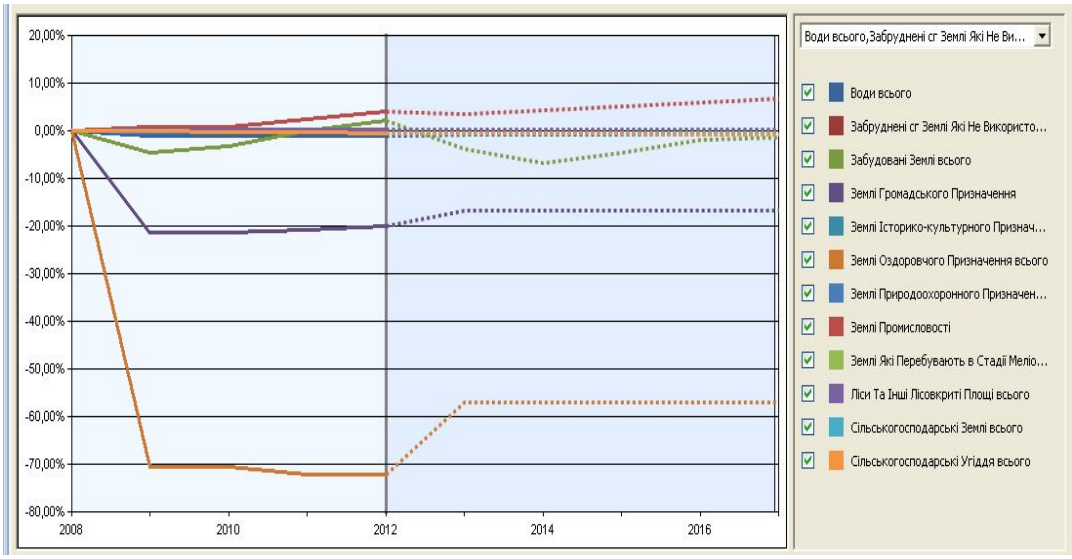
```
CREATE MINING MODEL [Forecasting_MIXED]
([Year] LONG KEY TIME,
[Water total]DOUBLE CONTINUOUS PREDICT,
[Polluted agricultural lands total]DOUBLE CONTINUOUS PREDICT,
[Built-up Lands total]DOUBLE CONTINUOUS PREDICT,
[Public lands]DOUBLE CONTINUOUS PREDICT,
[Lands Intended for Historical and Cultural purposes]DOUBLE CONTINUOUS PREDICT,
[Recreation lands]DOUBLE CONTINUOUS PREDICT,
[Lands lintended for Nature Protection purposes]DOUBLE CONTINUOUS PREDICT,
[Industry lands]DOUBLE CONTINUOUS PREDICT,
[Lands in the course of ameliorative construction and soil fertility restoring] DOUBLE CONTINUOUS PREDICT,
[Forests and other forested areas total]  DOUBLE CONTINUOUS PREDICT,
[Total area of agricultural lands]  DOUBLE CONTINUOUS PREDICT,
[Area of agricultural holdings total]  DOUBLE CONTINUOUS PREDICT)
USING Microsoft_Time_Series (AUTO_DETECT_PERIODICITY = 0.8, FORECAST_METHOD = 'MIXED')
WITH DRILLTHROUGH
```

Forecasting_MIXED model name must be entered in the first request line. Analysis Services will automatically form basic structure name by adding "structure" line to the model name which prevents from non-coincidence of structure and model names. Next code line is where key column of data mining model must be set, which in case with temporal series model will determine a unique tme phase for reference data.The time phase is defined with the help of key words KEY TIME after column name and data types. If

the temporal series model has a separate serier key, it will be defined with the help of the key word KEY. Therefore, "Year" column contain values of temporal phases used to define value sequence order. The following code lines are used to define determination of columns in the forecasted model. Parameter of the algorithm AUTO _ DETECT _ PERIODICITY = 0.8 shows that the algorithm must elicit cycles in the data. Setting a value close to 1 makes it possible to elicit a lot of patterns but may slow down the processing. Parameter of the algorithm FORECAST _ METHOD sets the data mining algorithm ARTXP or ARIMA, or a combination of these two algorithms.

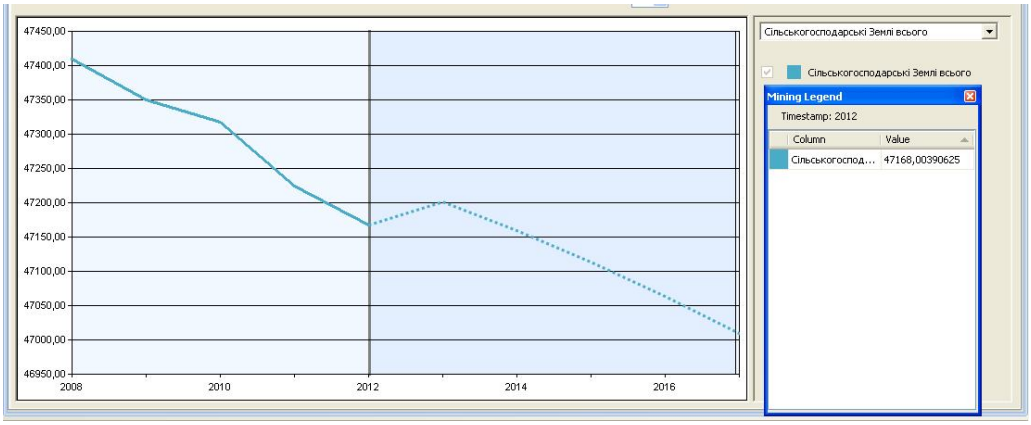
Key word WITH DRILLTHROUGH enables to drill through a detailed statistics in reference data after the model creation has been completed.

A process of forecasting values construction is divided in two phases – the first one is a study of data mining means based on a selected time series model, second – is a formation of the forecasting results by applying of the model studied. See graphical presentation of the forecasting results on pictures 1-4.



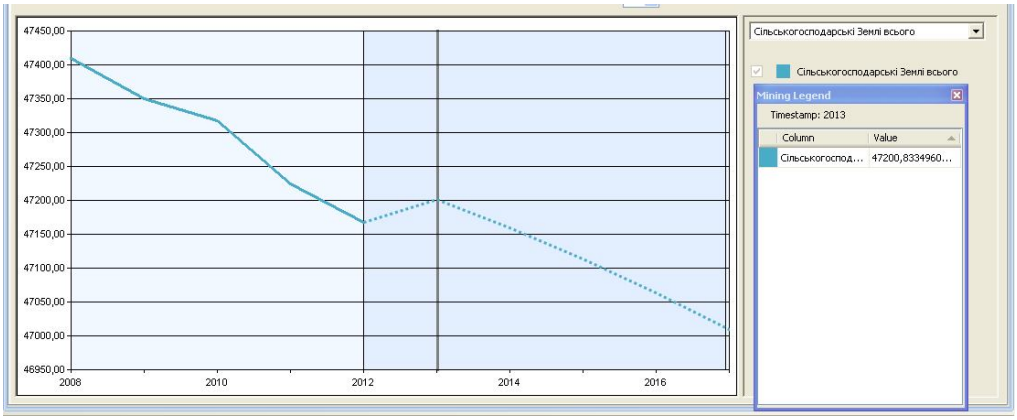
Picture 1. Page of the results of model applied to the land cadastre data

The diagram (Picture 1) shows both the past and the future. A part of graphic which is in charge of forecasted values is marked with a darker background. Each forecasted values is marked with the other colour. For the agricultural lands there are diagrams built with a cursor positioned at 2012, 2013, 2018 years, respectively (Picture 2, 3, 4.). When selecting a certain year, the legend shows a corresponding value of agricultural land area.

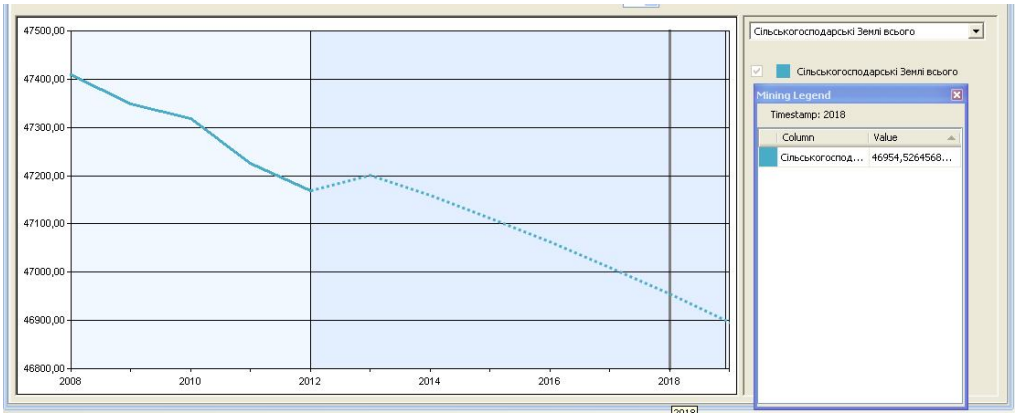


Picture 2. Diagram if change of agricultural lands, cursor positioned at the year 2012

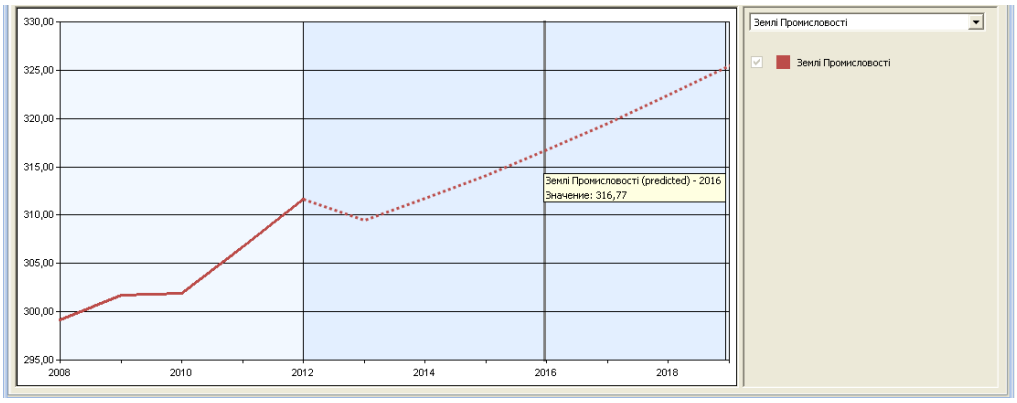
As we can see, values of agricultural lands area in 2012 (Picture 2) is 4,768.003 ha, and the forecasted area values in 2014 is 47,200.83 ha (Picture 3), in 2018 it is 46,954.52 ha (Picture4). The diagram (Picture6) shows that a partial reduction of industry land areas will take place in 2013, but they will grow during the subsequent periods.



Picture 3. Diagram if change of agricultural lands, cursor positioned at the year 2013



Picture 4. Diagram if change of agricultural lands, cursor positioned at the year 2018



Picture 6. Diagram of change of industry lands, cursor positioned at the year 2016

In order to obtain numeric results of the forecasted values, the following request was made:

```
SELECT
PredictTimeSeries([Forecasting_MIXED].[Water total], 6),
PredictTimeSeries([Forecasting_MIXED].[Polluted agricultural lands total], 6),
PredictTimeSeries([Forecasting_MIXED].[Built-up Lands total], 6),
PredictTimeSeries([Forecasting_MIXED].[Public lands], 6),
PredictTimeSeries([Forecasting_MIXED].[Lands intended for historical and cultural purposes], 6),
PredictTimeSeries([Forecasting_MIXED].[Recreation lands], 6),
PredictTimeSeries([Forecasting_MIXED].[Lands intended for nature protection purposes], 6),
PredictTimeSeries([Forecasting_MIXED].[Industry lands],6),
PredictTimeSeries([Forecasting_MIXED].[Forests and other forested areas total], 6),
PredictTimeSeries([Forecasting_MIXED].[Total area of agricultural lands], 6)
PredictTimeSeries([Forecasting_MIXED].[Area of agricultural holdings total], 6)
FROM [Forecasting_MIXED]
```

Function PredictTimeSeries in each forecasted line is implemented for preset fields. Numbers following the names of forecasted attributes show number of time steps required for the forecasting. See the request results in the Table 1.

Table 1.

Forecasted values of the land use.

Land resources	Forecasted values	
	Year	Values
Water total	2013	2030,67204589844
	2014	2030,67204589844
	2015	2030,67204589844
	2016	2030,67204589844
	2017	2030,67204589844
	2018	2030,67204589844
Polluted agricultural lands (total)	2013	0
	2014	0
	2015	0
	2016	0
	2017	0
	2018	0
Built-up Lands total	2013	2960,33473463929
	2014	2873,26946803778
	2015	2934,96439428538
	2016	3023,45628791804
	2017	3041,52843721532
	2018	3045,18295326832
Public lands	2013	350,295520019531
	2014	350,295520019531
	2015	350,295520019531
	2016	350,295520019531
	2017	350,295520019531
	2018	350,295520019531
Lands intended for historical and cultural purposes	2013	1,16999995708466
	2014	1,16999995708466
	2015	1,16999995708466
	2016	1,16999995708466
	2017	1,16999995708466
	2018	1,16999995708466
Recreation lands	2013	24,3473003387451
	2014	24,3473003387451
	2015	24,3473003387451
	2016	24,3473003387451
	2017	24,3473003387451
	2018	24,3473003387451
Lands intended for nature protection purposes	2013	3392,28803710938
	2014	3392,28803710938
	2015	3392,28803710938
	2016	3392,28803710938
	2017	3392,28803710938
	2018	3392,28803710938

Table 1.(Continuation)

Forecasted values of the land use.

Industry lands	2013	309,506332397461
	2014	311,716677770516
	2015	314,147994091302
	2016	316,766467991486
	2017	319,542720543041
	2018	322,451269483727
Forests and other forested areas (total)	2013	24807,279296875
	2014	24807,279296875
	2015	24807,279296875
	2016	24807,279296875
	2017	24807,279296875
	2018	24807,279296875
Total area of agricultural lands	2013	47200,8334960938
	2014	47159,0185515904
	2015	47112,8627400318
	2016	47063,0257777211
	2017	47010,0811308795
	2018	46954,5264568717
Area of agricultural holdings (total)	2013	46183,0568359375
	2014	46145,1242591658
	2015	46103,2522098726
	2016	46058,0393566741
	2017	46010,00610202
	2018	45959,6040567214

The analysis of the results received shows no changes of polluted agricultural lands, public, historical and cultural, recreation, nature protection purposes lands, land occupied with forests and water, increase of industry land area and reduction of agricultural lands and agricultural holdings.

Hence, the further land use requires due attention paid to the increase of the land use efficiency by applying efforts like land transformation.

Conclusions

Described herein data mining method based on a forecasting of time series for land cadastre makes it possible to make data analysis and forecast future values of the land use in Stryi district Lviv region.

A defect of such forecasting method is that is uses no independent variables and is fully based on the forecasted series history.

1. Fayyad U.M. Advances in Knowledge Discovery and Data Mining [Text] / U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. - Menlo Park, Calif.: AAAI/MIT Press, 1996. 2. Барсегян А.А. Методы и модели анализа данных OLAP и DataMining [Текст] / А. А Барсегян., М.С. Куприянов, В. В. Степаненко, И. И.Холод. – СПб: БХВ-Петербург. – 2008. 3. Бокс Дж., Дженкинс Г. Анализ временных рядов, прогноз и управление: Перевод с англ. [Текст] //Под ред.В.Ф,Писаренкою. «Мир», М., 1974. 4. Time Series Prediction DMX Tutorial [Електронний ресурс]. – Режим доступу: http:// http://technet.microsoft.com/en-us/library/cc879270.aspx.