Intelligent technologies for filling gaps in the data tables of audit objects

IT for Practice 2022

VSB-TU Ostrava, 13-14 October 2022

Vyacheslav Chaplyha
Department of Information Technology
Lviv National Environmental University

Volodymyr Chaplyha
Department of Accounting and Auditing
Ivan Franko National University of Lviv

Nataliya Abashina
Department of Ophtalmolodgy
Danylo Halytsky Lviv National Medical University

AGENDA

- *Introduction*
- *Statement of the problem*
- *Statement of the neuron-like structure of the auto-associative type*
- *IT for filling gaps in the data based on the use of auto-associative neural network*
- *Software implementation of a neural network for filling gaps in data tables*
- *Conclusions and recommendations*

Introduction

- *Problems of filling gaps in data tables arise in many cases of analysis, classification and forecasting of information objects of various origins and nature, functioning under conditions of uncertainty, in particular economic, social, political, financial and audit.*
- *As is known [1], the problem of filling gaps in data tables belongs to the class of incorrect, that is, those that do not have a single solution. To solve such problems, a number of approaches are used, the specific choice of which will depend on both the features of the table data and the subjective preferences of the researcher.*
- *As a result, in many cases one can only hope to obtain a "plausible" filling of gaps.*

Statement of the problem

Traditionally, the following methods are used as filling methods:

- *averaging,*
- *zeroing,*
- *extraction of missing values during mathematical processing [1], as well as separate regression models.*

Obviously, the highest quality of results should be expected in the case of regression methods, which, however, is achieved only under the condition of mutual correlation of the table data, as well as the implementation of a number of restrictions on the amount of missing data. In particular, it is required to observe certain ratios between the number of rows with and without gaps, etc.

Statement of the problem

In case of interdependence of table elements, only one of the first three methods listed above can be used. The quality of such filling in most cases is too low, but there is no alternative in such a situation. Significantly expanded opportunities open up for table variants in which certain correlations between elements can be traced, in particular

- - there are interdependencies between elements of rows of each separate column; such dependencies take place for time sequences, or in the presence of a separate column (columns), which is a marker of each row of the table;
- - there are interdependencies between elements by columns for each individual row; this option is more universal, since it also implies the option of dependencies between individual rows.

Solution of the problem

The use of neural-like structures of the Geometric Transformation Machine (GTM) [2] provides a universal approach suitable for various variants of data tables, with the presence of different types of relationships between their elements.

The method is based on the spatial interpretation of the table of interdependent data as a body, which is a geometric place of points - table rows in coordinates, each of which corresponds to one column of the table.

That is, each row of the table is considered as a point in the space of realizations, the dimension of which is determined by the number of columns of the table, and each column represents one of the coordinates of the body.

Statement of the neuron-like structure of the auto-associative type

The use of neural-like structures of the Geometric Transformation Machine (GTM) [2] provides a universal approach suitable for various variants of data tables, with the presence of different types of relationships between their elements.

The method is based on the spatial interpretation of the table of interdependent data as a body, which is a geometric place of points - table rows in coordinates, each of which corresponds to one column of the table.

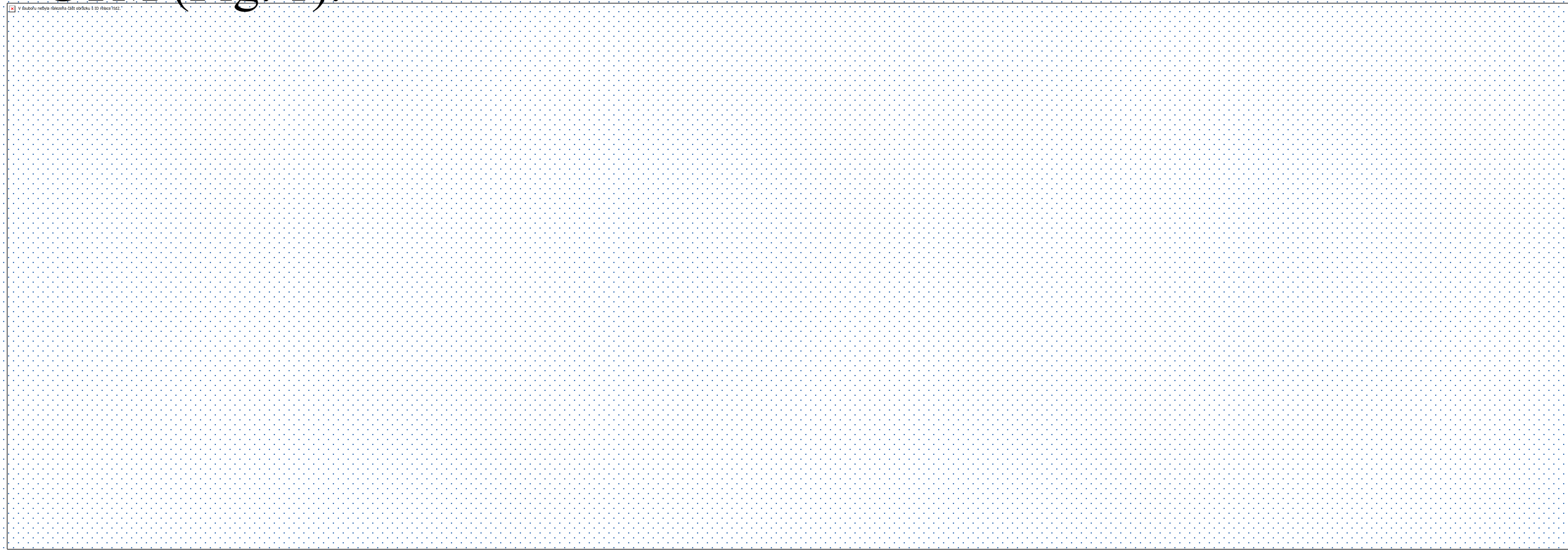
That is, each row of the table is considered as a point in the space of realizations, the dimension of which is determined by the number of columns of the table, and each column represents one of the coordinates of the body.

Statement of the neuron-like structure of the auto-associative type

The method of filling gaps based on the GTM is based on the principle of constructing a body that can be partially or completely represented only by projections of points on certain coordinate hyperplanes and finding all the coordinates of points by their projections, based on the belonging of each point to the constructed body. This method is superior in quality to the existing ones, primarily because it ensures that the full information represented by the given components of the elements of the table a is taken into account, therefore, it allows you to get more accurate representations, if it is generally possible for a given set and structure of data.

IT for filling gaps in the data based

- Let us consider the basics of the functioning of the gap filling system based on the neuron-like structure of the auto-associative type of the GTM (Fig. 1).



IT for filling gaps in the data based

Information technology of filling gaps in the data based on the use of auto-associative neural network GMT involves the following steps of information transformation:

- 1. training of the neural network on the basis of training vectors - rows of data in which gaps are replaced by average values;
- 2. translation of training vectors-strings through the trained neural network to the output;
- 3. replacing the initially missed vector components with the values obtained at the outputs of the neural network;
- 4. exit from the loop if the specified threshold of changes between two consecutive transformations is reached
- 5. retraining of the neural network;
- 6. transition to step 2.

IT for filling gaps in the data based

This approach is based on the unique property of GMT neural networks to quickly retrain, is much more accurate than other methods of filling gaps, is universal, as it is suitable for preliminary data analysis, and for making replacements in tables of different structure and content, does not require special knowledge from the user.

Software implementation of a neural network for filling gaps in data tables.

The Expleo XL software product [2] implements a nonlinear method that does not impose additional requirements on the processed data, and therefore can be used in almost all cases. The case when each row of data contains gaps in different columns is also acceptable.

Software implementation of a neural network for filling gaps in data tables

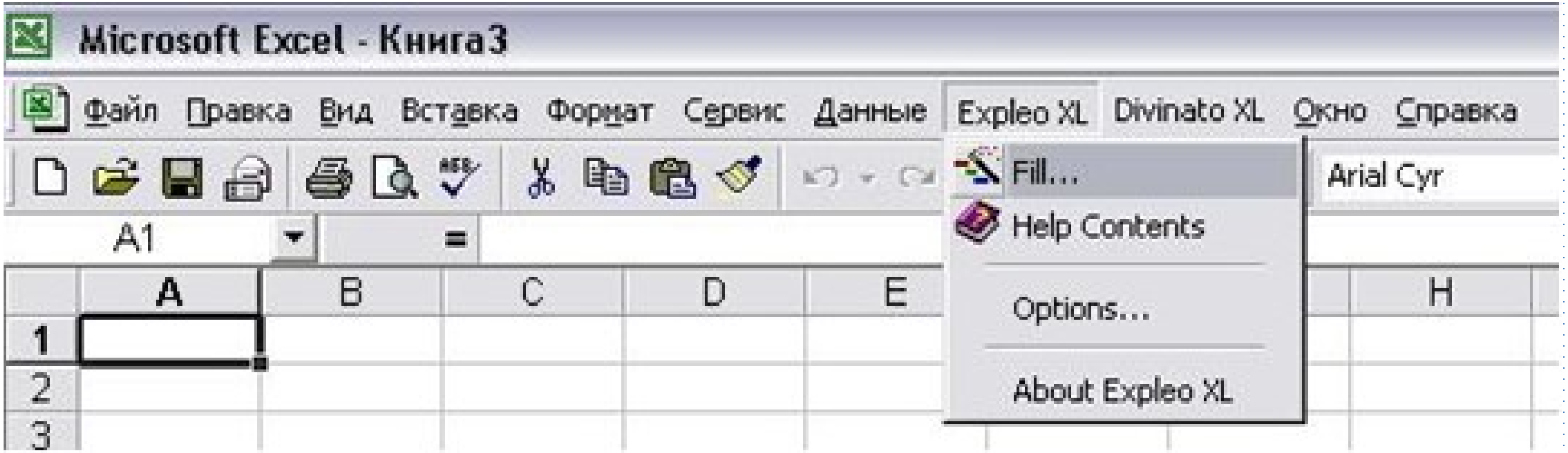
This approach is based on the unique property of GMT neural networks to quickly retrain, is much more accurate than other methods of filling gaps, is universal, as it is suitable for preliminary data analysis, and for making replacements in tables of different structure and content, does not require special knowledge from the user.

Software implementation of a neural network for filling gaps in data tables.

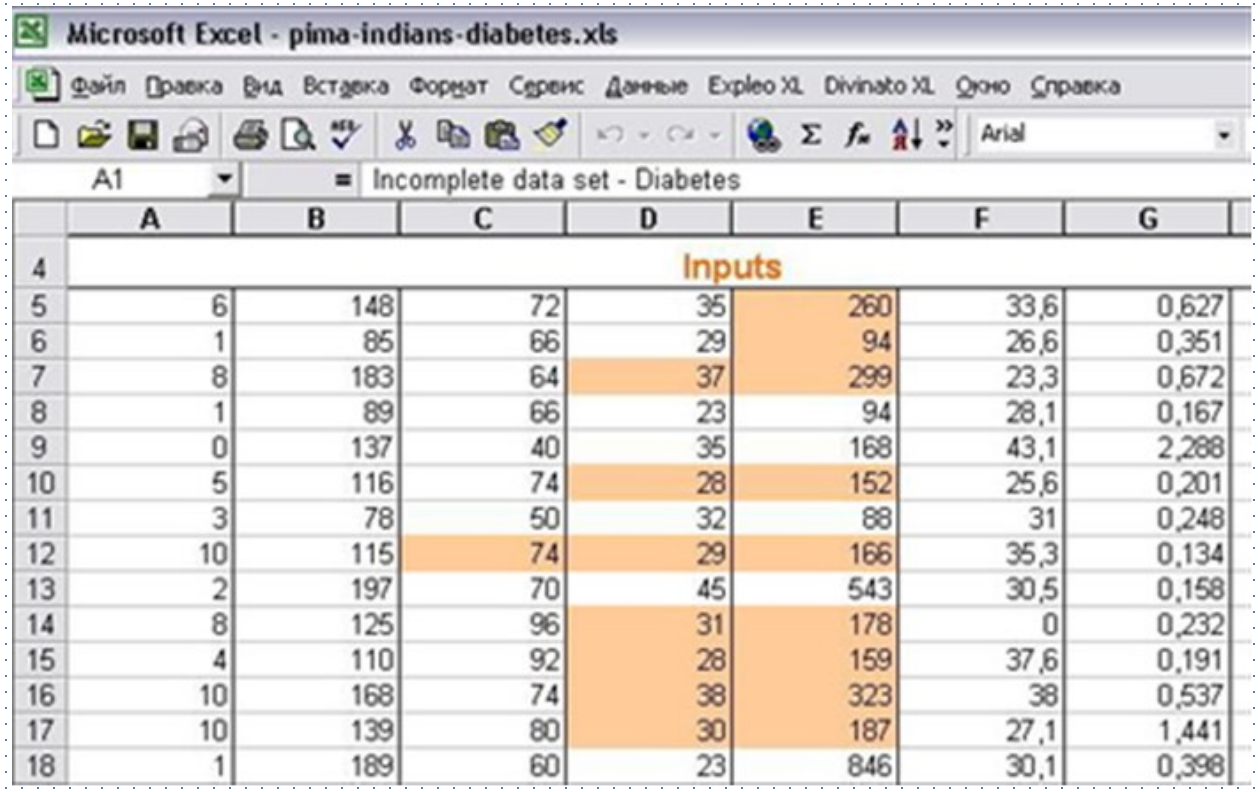
The Expleo XL software product [2] implements a nonlinear method that does not impose additional requirements on the processed data, and therefore can be used in almost all cases. The case when each row of data contains gaps in different columns is also acceptable.

Software implementation of a neural network for filling gaps in data tables

Expleo XL integrates into the MS Excel menu.



Software implementation of a neural network for filling gaps in data tables



	A	B	C	D	E	F	G
4	Inputs						
5	6	148	72	35	260	33,6	0,627
6	1	85	66	29	94	26,6	0,351
7	8	183	64	37	299	23,3	0,672
8	1	89	66	23	94	28,1	0,167
9	0	137	40	35	168	43,1	2,288
10	5	116	74	28	152	25,6	0,201
11	3	78	50	32	88	31	0,248
12	10	115	74	29	166	35,3	0,134
13	2	197	70	45	543	30,5	0,158
14	8	125	96	31	178	0	0,232
15	4	110	92	28	159	37,6	0,191
16	10	168	74	38	323	38	0,537
17	10	139	80	30	187	27,1	1,441
18	1	189	60	23	846	30,1	0,398

Expleo XL automatically distinguishes between data represented by text (classes) and data represented by integers and floating point numbers. If some columns contain classes represented by numbers, you need to manually specify the type of these columns.

In addition to the usual information in tables, this method can also fill in gaps in time series. The Time Series mode is activated by pressing the Time Series button.
Next, another setting parameter appears - Period, which determines the periodicity of the data. The periodicity depends on the nature of the data: for example, when the table contains monthly reports, you should probably try to use the period 12 (covers 1 year).

Conclusion

The approach described above is universal, quite simple and provides, in general, a noticeably higher quality of filling gaps in the data.

For available data of a larger volume, the accuracy of filling increases; this also applies to the number of columns in the data table.

Completely empty columns will not be filled; completely empty rows will be filled only in the time sequence mode; in this case, it should be taken into account that the number of consecutive empty cells cannot exceed the length of the period.

Thank Your for attention!

Vyacheslav Chaplyha,

Volodymyr Chaplyha,

Nataliya Abashina

4vyach@gmail.com