Recovery gaps in experimental data

Roman Kaminskyi, Nataliia Kunanets [0000-0003-3007-2462],

Volodymyr Pasichnyk [0000-0003-3007-2462],

Antonii Rzheuskyi [0000-0001-8711-4163],

Andrii Khudyi

[1] Information Systems and Networks Department,
Lviv Polytechnic National University,
Stpan Bandera street, 32a, 79013, Lviv, Ukraine

Roman.M.Kaminskyi@lpnu.ua;
nek.lviv@gmail.com;
Volodymyr.V.Pasichnyk@lpnu.ua;
antonii.v.rzheuskyi@lpnu.ua;
Khudyy@ukr.net

Abstract. In advanced information technology of statistical analysis, often data for which there are no properties, parameters, characteristics and their values is found. In this situation, the actual becomes the problem of recovering missing data. It's almost impossible to set a value which is missed, but there is a large number of simple and more complex methods for replacing these values. This study describes the characteristics of some methods of filling gaps and examples of their application to the tables of data and time series.

Keywords: exclusion method, replacement methods, filling gaps with the average value, filling gaps with median, method of closest neighbors, filling data model with value, filling gaps without matching, regression analysis, maximum likelihood method, EM algorithm, ZET algorithm.

1 Introduction

Database tables and time series are widely used to describe various static and dynamic systems. However, in analyzing the results of studies of different systems there are often situations when the values of certain data are missing. In such cases, we have data with gaps, which greatly complicates the processing of data, since estimations of statistical characteristics have displacement. Recovering of missing data is a primary procedure and involves not only the application of certain methods for recovery of missing data, but also the knowledge of their nature.

The problem of missing data values is very relevant in sociology, image recognition, cluster analysis, and so on. Most often, it also occurs during the identification of time

series, when a priori information about the value of the parameters is incomplete. Objective reasons for the occurrence of gaps are often: failure of equipment for measurement and registration, the emergence of obstacles in the monitoring, overlay on the observed attributeal of interference, etc., and subjective reasons – inactivity in data registration, inability to get full and accurate information through the influence of the fuzzy qualitative situation, psychological aspects and attributes of memory, as well as the delayed sensory-motor reaction.

2 Causes and problems of missing data recovery

The presence of missing values in data, attribute ificantly limits the possibility of using the required methods of processing. In this situation, an important parameter is the frequency of occurrence of gaps and the existence of certain regularities in them.
In complex systems often two types of data are formulated: tables and time series.
The tables data are presented as follows: the rows are the characteristics (parameters) of the studied objects (states, computers, methods, etc.), and columns are the value of a specific characteristic for each object and, mostly, they are numerical values generated by predetermined scale.
Time series characterize the dynamics of object as a change of some specific defining indicator, the value of which most adequately reflects the behavior of the object.
Situations requiring decision-making, generate the following needs for the recovery of missing data:

1. Filling all gaps in tables or time series.
2. Filling only some gaps.
3. Filling the gaps based on information contained in the table.
4. Filling each subsequent gap, based on the analysis of the initial information and obtained as a result of predicted values, taking into account the trends for previously filled gaps.

Existing methods of filling the gaps substantially different and provide different needs of such recovery. At the same time, the quality of data recovery is an important point in filling in the missing data.
It should be taken into account that the values of the characteristics in the table or the values of the time series levels are obtained in one way or another by data that may contain gaps, but only the tables themselves or the time series contain the most information about these data. In the case of large amount of data and small number of gaps, characteristics of the obtained data will vary slightly from the true values for their entire population (full data availability). In this case, to find a replacement for the missing value is not very difficult, since you can set the nature of the data (at least statistically). Otherwise, it is necessary to use several methods and choose the best to achieve a determined quality criterion, for example, for the characteristics of descriptive statistics, the mean square deviation from the trend, its various models.
R. Little [1] considered it expedient to process incomplete data, in case of missing time series levels (considering existing levels as elements of the sample), use the

following methods, which consist in determining the primary statistical characteristics of the studied sample:

1. without gaps;
2. when filling missing values with zeros;
3. when filling missing values with average values;
4. when filling missing values with the indicators of numerical characteristics of the distribution levels – mod, median or its quartiles;
5. when filling the missing levels quasi-random numbers, distributed by the normal law, the average value and the mean square deviation of the levels coincide with the corresponding characteristics of the initial series with the gaps.

3 Methods for filling the spaces in the data tables

The practice of working with databases gives reason to assert that there is a high likelihood of a large number of tabs in the table, in particular in the "object-attribute" Table 1. The attributes of objects in this table are certain physical parameters, each of which has its dimension. Therefore, to fill the spaces given in the table with an empty space character, use the comparison methods with the analogs. Statistical analysis of the values of different attributes of one object is inadmissible because it leads to errors and loss of confidence in the data. Recovery of missing data in tables makes sense only in the case of several gaps.

Table. 1. "Object-attribute" with missing values.

Objects	**Attributes**					
	Height	**Length**	**Width**	. . .	**Weight**	**Volume**
Object Y_1	h_1	d_1	l_1	⋯	m_1	v_1
Object Y_2	h_2	$\varnothing$	l_2	⋯	m_2	v_2
. . .	⋯	⋯	⋯	⋯	⋯	⋯
Object Y_{n-1}	h_{n-1}	d_{n-1}	$\varnothing$	⋯	m_{n-1}	$\varnothing$
Object Y_n	h_n	$\varnothing$	l_n	⋯	m_n	$\varnothing$

The most common methods that can recover missing data without losing their reliability are the following:

***Exclusion method*.** This method is used to recover gaps in tables and is implemented by eliminating lines with the presence of gaps. In general, this reduces the informality of the tables and the loss of objectivity in the analysis of the situation, based on the processing of the data thus obtained, reducing the adequacy of the constructed models.

Replacement methods are diverse, we consider only some of them, which in our opinion, expedient to use for recovering of gaps in the data obtained during the scientific experiment.

Filling gaps with the average value of the data presented in the column. Missing data are generated by calculating the average value available in the table of values of given attribute, assuming that the objects are equivalent. At the same time, the application of this method is possible only if there is no data in random order MAR (missing at random) [2], when gaps are random variables. The disadvantage of the method is distorting the distribution of data and dispersion decreasing of initial data.

Filling gaps with median.[3] The median is the most stable characteristic of the sample, since in any transformation it remains unchanged. It can be successfully used for a table. However, in the case of a small number of objects and several gaps the median value can vary quite attributeificantly, so the median value must be determined after each completed gap, and the gaps fill in a random way.

Method of closest neighbors. The method is based on the assumption that the missing value is close to the filled values of the rows, neighboring with the row with missing value. To fill the value of the missing attribute, the values of all relevant attributes of these neighboring rows are averaged. In this case weight coefficients are used, which are inversely proportional to the distance between the cell with the gap and the cell with available value of the given attribute. For a large number of gaps, this method is ineffective.

Filling data model with value. This method is building the model of values of the given attribute. For filling take the value of this model, which corresponds to a gap.

Filling gaps without matching. The pass is filled with the constant value from an external source, for example, a value of the previous observation from the same study.

Regression analysis. [4] For application of this method it is necessary to comply with the requirement of data compliance to condition MAR, as well as requirements related to the implementation of the prerequisites of regression analysis. The disadvantages of this method include the dependence of quality of gaps recovery from choice of regression model. These methods belong to the category of simple techniques and are usually performed in the pre-processing of data and preparation for analysis. In addition to these methods, used others, classified as complex. In turn, this category is divided into two subcategories – global methods and local methods. The following are the global ones.

Maximum likelihood method. **[**5**]** The basis of this method is finding the maximum values of mathematical expectations, which are target variables for each missing value, using the existing observations. Its application is complicated by the large number of missing values.

EM-algorithm. **[**6**]** This algorithm is implemented in two stages. In the first stage, which is called stage **E**, using full or partly incomplete data, determine the conditional mathematical expectations, which fill each proxy value. After filling all the missing values, determine the average, dispersion, correlative and covariance indicators. In the second stage, which is called stage **M**, they achieve the maximum matching of the substituted found values so that data structure with the filled variables matches data structure of complete observations.

Algorithm ZET. It refers to local fill-in gaps algorithms. These algorithms mainly take into account the dependencies, for that part of the data containing the gap, that is, in some area of gap. In identifying the dependence for this algorithm, all the rows and columns of the output data field are involved. Local algorithms have high efficiency in comparison with other known algorithms for filling the gaps. For real-world tasks, different their modifications are used.

In the table, when data values are missed, that is, when some of the objects present are incomplete sets of attributes, often act in the following way.

1. To fill the missing values, select similar objects with complete sets of attributes, that is, objects that do not have missing values of attributes.
2. On subset of these objects, different missing values are simulated, typical for this type of data.
3. Data recovery is carried out with various methods.
4. Next, determine which of the methods provides the best match for the replacement of missing values calculated, within a given criterion.
5. This method is used to recovery the really missing values of the objects attributes of the given set.

Example. To illustrate the methods we consider a hypothetical example of filling in the table of missing values in the environment of the table processor Ms Excel. Objects in the table are characterized with only one attribute – the values of the indicator. As such vector of attributes a sequence of random numbers C3: C32 with steady distribution within the values of attributes is used $x_i \in [1, 10]$ as shown in Fig. 1. This sequence includes $n = 30$ values and is practically the least representative, and therefore conclusions based on it can be considered reliable.

	number object	value original	missing	average	weighted average	median
1	1	9,096	9,096	9,096	9,096	9,096
2	2	3,407	3,407	3,407	3,407	3,407
3	3	9,567	9,567	9,567	9,567	9,567
4	4	3,031	3,031	3,031	3,031	3,031
5	5	8,652		6,156	2,859	5,770
6	6	8,712	8,712	8,712	8,712	8,712
7	7	1,371	1,371	1,371	1,371	1,371
8	8	8,556	8,556	8,556	8,556	8,556
9	9	9,977	9,977	9,977	9,977	9,977
10	10	5,895	5,895	5,895	5,895	5,895
11	11	3,758		6,156	3,838	5,770
12	12	9,332	9,332	9,332	9,332	9,332
13	13	5,573	5,573	5,573	5,573	5,573
14	14	5,101	5,101	5,101	5,101	5,101
15	15	9,855	9,855	9,855	9,855	9,855
16	16	4,809	4,809	4,809	4,809	4,809
17	17	2,122	2,122	2,122	2,122	2,122
18	18	2,687		6,156	2,339	5,770
19	19	3,872	3,872	3,872	3,872	3,872
20	20	9,990	9,990	9,990	9,990	9,990
21	21	2,187	2,187	2,187	2,187	2,187
22	22	5,756		6,156	2,972	5,770
23	23	5,645	5,645	5,645	5,645	5,645
24	24	8,462	8,462	8,462	8,462	8,462
25	25	2,236	2,236	2,236	2,236	2,236
26	26	7,692	7,692	7,692	7,692	7,692
27	27	1,127	1,127	1,127	1,127	1,127
28	28	4,522	4,522	4,522	4,522	4,522
29	29	8,411	8,411	8,411	8,411	8,411
30	30	9,499	9,499	9,499	9,499	9,499
	Descriptive Statistics					
Mean		6,030	6,156	6,156	5,735	6,104
Standard Error		0,544	0,600	0,519	0,557	0,519
Median		5,700	5,770	6,156	5,337	5,770
Mode		#N/A	#N/A	6,156	#N/A	5,770
Standard Deviation		2,981	3,059	2,840	3,049	2,844
Sample Variance		8,885	9,359	8,068	9,297	8,086
Kurtosis		-1,487	-1,484	-1,191	-1,582	-1,217
Skewness		-0,117	-0,214	-0,228	0,085	-0,169
Range		8,863	8,863	8,863	8,863	8,863
Minimum		1,127	1,127	1,127	1,127	1,127
Maximum		9,990	9,990	9,990	9,990	9,990
Sum		180,897	160,044	184,666	172,052	183,123
Count		30,000	26,000	30,000	30,000	30,000

Fig. 1. The results of filling missing values.

Let for several objects there is no given attribute, that is, in the relevant column of the table "object-attributes" there are missing values. To fill the gaps we use: the average value of the sample with gaps, the weighted average of this sample and the median value. To simulate the gaps remove from this sequence the following values C7, C13, C20, and C24. As a result, we will obtain the vector of attribute with gaps E3:E32.
The value of the average and median are determined using the procedure «Data → Data Analysis → Descriptive Statistics». The weighted average is determined by the following formula

$$x_i = 0.5(0.191 \cdot x_{i-2} + 0.309 \cdot x_{i-1} + 0.309 \cdot x_{i+1} + 0.191 \cdot x_{i+2} \quad (1)$$

where x_i – missing value.

Indicators of descriptive statistics for the recovered are shown below in Fig. 2. The values of the objects attributes have steady distribution, and therefore the main characteristics in descriptive statistics, which can be compared with the results of the application of one or another method are: arithmetic mean, median, standard deviation and sum (value of kurtosis, asymmetry and mod are not informative and incorrect). According to these indicators the relative error of the data with missing values and data with replacement of missing values in relation to the original data – values C3:C32. The values of relative error are presented in Table 2.

Table 2. The values of relative error.

Methods of filling missing values	Relative error of indicators of descriptive statistics in [%]			
	Average	Median	Standard deviation	Sum of values
With gaps	2.089	1.228	2.617	11.528
Average arithmetic	3.267	8.000	4.730	2.083
Weighted average	4.892	6.368	2.281	4.889
Median	1.227	1.228	4.596	1.231

In this example, the most suitable was the method of filling missing data with median value.

Remark. This example illustrates only a procedure, rather than solution to specific problem.

As the criterion for assessing the quality of the method is the relative error of indicator value for characteristics of descriptive statistics, then the smaller this value, the better is replacement of value with the specified method.

4 Methods of filling gaps in time series

Data representation of time series and their analysis is becoming increasingly popular in various scientific studies. Especially time series analysis is important for research of data streams in information systems and networks, in problems of modeling processes of different systems and phenomena, in predicting situations and dynamic of systems on the basis of monitoring of their state.

The main reason that causes the gaps in time series is the inability to obtain information at certain points of time. Besides, it may be a situation where the means of measurement, observation, registration are not configured, damaged, do not meet the measured values or have inappropriate limits of their measurements (discrepancy of scales of values, low sensitivity, require a considerable amount of time for measurement), data is recorded by unskilled personnel.

Characteristic for time series is that, depending on subject area, the nature of gaps has its own peculiarities. However, in the process of filling gaps, their nature is often ignored and one or more of the most accessible and simple methods are applied. View of time series with gaps is shown in Fig. 2.

Today there is no single methodology for recovering of missing values or processing missing data. The choice of the most appropriate method for filling gaps in each particular situation is often a rather complicated individual task, which can take much more time and efforts than data processing itself with recovered values.

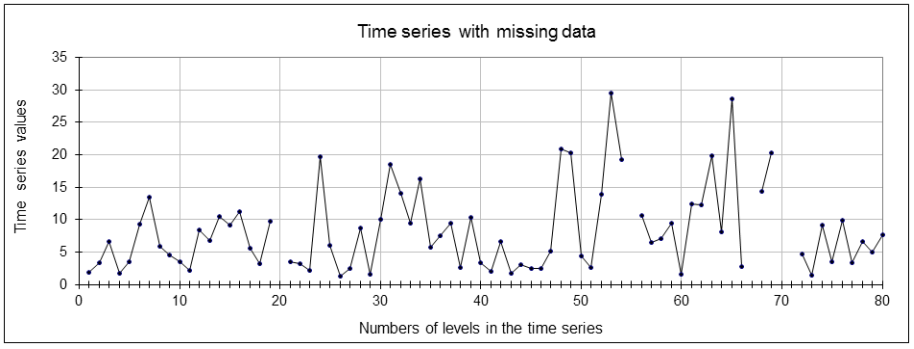

Fig. 2. Time series with gaps look like a sequence of individual fragments.

To fill the gaps in time series can be used different methods, but in each case the method of filling must be substantiated, and the results interpreted. Unlike tables in time series, the attributes are equal in time series, but they have the same nature, physical content, their values are measured in the same scale, they are dependent random variables with the same distribution and, most importantly, they are connected with ordered sequence of moments of time, in which their registration was made.

To fill missing values, levels in time series, we can distinguish two of these methods: use the values of individual statistics and use the values of the trend model.

To fill the missing levels are used the following statistics.

Filling the missing levels with average value. For time series, the gap is filled in or with general average for all values of the series, or selecting a certain interval inside which is the gap and the average is calculated for this interval for filling the gap. This method is easy to implement, and the mechanism of creation of gap can be ignored. The disadvantage of the method is distorting the distribution of data and reducing the dispersion of initial data.

Filling gap with median. The median is the most stable characteristic of the sample, since in any transformation it remains unchanged. It can be successfully used for the time series as well.

Filling with distribution mode. In the case when it is necessary to find value of missing level of time series, for sufficient amount of data, the value of mode is determined, which is used to recover the gap.

Remark. The lack of inflection point of envelope of variation series indicates that there is no mode for the distribution of time series levels.

Filling values of trend model. The essence of this method is that a trend model is built in the form of appropriate function, mainly nonlinear. Then the values of missing levels are taken from this model (function) in accordance with the numbers of these levels.

Example. For the time series shown in Fig. 3. the most appropriate method for filling missing levels is to calculate their values based on the model of its trend.

Let the output series have $n = 80$ levels, however, there are gaps, namely missing levels x_i there are levels for which $i = 20, 55, 67, 70, 71$.

Filling the missing levels is carried out in following way. A trend model is being constructed using as an approximation function, for example, a third-order polynomial. This choice is conditioned with the following consideration. Because the form of the trend is unknown, the selected approximation function should reflect growth, decline, certain changes that may have a tendency. The third degree polynomial is actually a cubic parabola, which has an inflection point, which means that it can "catch" existing, although quite common, changes in trends.

The procedure is that first we find the trend of the original series and approximate it with the third degree polynomial. Define the value of the first gap from trend equation and fill it the first missing value. Next, we approximate the series with the filled gap by the same polynomial and determine from obtained trend model (approximating function) value for the second gap. As a result of the completion of this procedure, the time series graph has the form shown in Fig. 3.

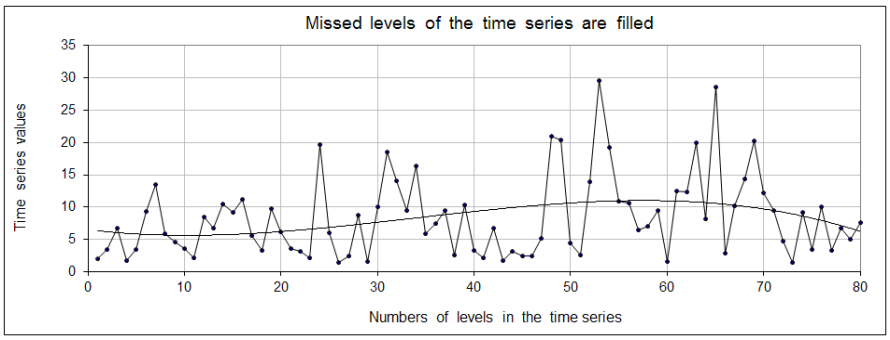

Fig. 3. Time series with filled gaps.

In fig. 3. thin line reflects the trend, approximated with polynomial of the third degree after filling in the missed fourth level.

The analysis of coefficients, approximating trend, polynomials showed that in this case, they are very weakly different from each other, and and almost all trend lines merge because the difference between them is much smaller than that for the given scale and pixel sizes can be submitted. For a more detailed analysis, we'll compare the changes of parameters of descriptive statistics for presented series, which are presented in Table 3. Based on the results of estimating the parameters of descriptive statistics, we can conclude, that almost every parameter "feels" a change of time series when replacing the missing value with a value determined with using the model. No changes were made within the limits of accuracy of "three decimal places" only mode, interval, minimum and maximum values. This can be explained by the fact that the replacement of the gaps obtained by values from the models, practically does not affect the distribution of time series levels. The change of data values of kurtosis and asymmetry of distribution can be considered insignificant as these parameters determine the features of shape of curve of the law of distribution and density functions, which are visually impossible to capture Interval, minimum and maximum values remained unchanged because the values of the model lie inside the interval – do not exceed the extreme values of levels. Thus, the problem of filling gaps in real tables or time series can be solved with different methods, however, it is always necessary to keep in mind the features of the source that generates these data, the situation at the time of data collection, as well as the requirements to use data with filled gaps**.**

Table 3. The changes of parameters of descriptive statistics.

Parameters of descriptive statistics	Time series with missing values	Filling in missing values				
		x_{20}	x_{55}	x_{67}	x_{70}	x_{71}
Average	**8,235**	**8,209**	**8,243**	**8,268**	**8,318**	**8,332**
Standard error	**0,739**	**0,730**	**0,721**	**0,712**	**0,705**	**0,696**
Median	**6,703**	**6,688**	**6,703**	**6,703**	**6,703**	**6,742**
Mode	**6,703**	**6,703**	**6,703**	**6,703**	**6,703**	**6,703**
Standard deviation	**6,401**	**6,362**	**6,327**	**6,290**	**6,265**	**6,227**
Dispersion of levels	**40,967**	**40,474**	**40,03**	**39,56**	**39,251**	**38,769**
Kurtosis	**1,660**	**1,733**	**1,740**	**1,766**	**1,735**	**1,777**
Asymmetry	**1,349**	**1,367**	**1,355**	**1,349**	**1,326**	**1,326**
Interval (scope)	**28,187**	**28,187**	**28,18**	**28,18**	**28,18**	**28,18**
Minimum	**1,360**	**1,360**	**1,360**	**1,360**	**1,360**	**1,360**
Maximum	**29,547**	**29,547**	**29,54**	**29,54**	**29,547**	**29,547**
Sum	**617,625**	**623,84**	**634,7**	**644,9**	**657,10**	**666,52**
Number of levels	**75,000**	**76,000**	**77,00**	**78,00**	**79,000**	**80,000**

Conclusions

This researh shows that even using quite simple methods of filling missing values in tables and time series, on condition of representative amount of data, you can get quite good results in replacement of the missing data.

The use of indicators of descriptive statistics as a criterion for evaluating the replacement of missing values is completely correct and quite sufficient for most real situations. Obviously, the random data considered in the examples can be attributed to stationary random sequences, at least in relation to their average value and dispersion. However, in the case of significant no-nlinearities and a relatively small amount of missing values, the basic method is the construction of empirical model and matching of its properties in the framework of the set task.

References

1. Little, R., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons, Inc. (1987).
2. Abnane, I., Abran, A., Idri, A.: Missing data techniques in analogy-based software development effort estimation. Journal of Systems and Software, 117, 595–611 (2016).
3. Valencia, Pedro L., Astudillo-Castro, Carolina, Gajardo, Diego, Flores, Sebastián: Calculation of statistic estimates of kinetic parameters from substrate uncompettive inhibition equation using the median method. Data in Brief 11, 567–571 (2017).
4. Mika Sato-Ilic: Knowledge-based Comparable Predicted Values in Regression Analysis. Procedia Computer Science 114, 216–223 (2017).
5. Dong-Qing Wang, Zhen Zhang, Jin-Yun Yuan: Maximum likelihood estimation method for dual-rate Hammerstein systems. International Journal of Control, Automation and Systems 15(2), 698–705 (2017).
6. Balakrishnan, Sivaraman; Wainwright, Martin J.; Yu, Bin: Statistical guarantees for the EM algorithm: From population to sample-based analysis. Ann. Statist. 45 (1) 77–120 (2017).