Evaluation of Accuracy of the Estimation Methods for Replacing Missing Values for Time Series Variables. Using the statistical packages software (SPSS &MINITAB1

Dr. Marie M. Mahmoud

Dept. of Statistics & Mathematics
Facially of Camm.. Tanta University

ABSTRACT

Missing values define specified data values as user-missing. It is often useful to know why information* is *missing. For example, you might want to distinguish between data missing because a respondent refused to answer and data missing because the question didn't apply to that respondent. Data values specified as user-missing are flagged for special treatment and are excluded from most calculations. You can enter up to three discrete (individual) missing values, a range of missing values, or a range plus one discrete value. Ranges can be specified only for numeric variables. In this paper, we aim at to determine the estimation methods using SPSS and MINI TAB packages for replacing missing values for time series variables, which minimize the total prediction error (dire(forecasting error and indirect forecasting error or replacing processing error), according to three Measures of Accuracy of the fated model:)1)' can Absolute Percentage Error (MAPE), Mean Absolute Deviate, a (MAD), and Mean Squared Deviation (MSD) in Time Series. In this toper, we consider the difference between the continuous and discrea distributions, also the difference between the symmetric and assynt, ferric distributions, and so the difference between the moderate and hr toy tailed distributions.

Keywords: *Time Series Data, Missing Values, Series mean, Mean& Median of nearby points, linear interpolation, linear trend at point, Measures of Accuracy, and Forecasting Error.*

(1). INTRODUCTION

Reducing the estimator bias is very important in statistical analysis. In this paper, five methods for replacing missing values for time series variables arc applied, and we aim at to determine the best estimation method, according to: 1- The probability distribution of the variable:
2- Three measures of accuracy of the fitted model.

To set the scene. **I first need to briefly describe what I mean** by:

Time **Series Data:** Corresponds to the sequence of values for a single variable in ordinary data analysis. Each case (row) in the data represents an observation at a different time. The observations must be taken at equally spaced time intervals.

Creating Time Series: Creating Time Series means creates new variables bated on functions of existing numeric time series variables. These transformed values are useful in many time series analysis procedures.

User-Mining Values: Values you have specified as missing, using Define Variable in the Data menu. You can specify individual missing values for numeric or string variables or a range of missing values for numeric variables. See also system-missing values.

System-Missing Values: Values assigned by the program when values in your data are undefined according to the format type you have specified, when a numeric field is blank, or when a value resulting from a transformation command is undefined. Numeric system-missing values are displayed as periods. String variables cannot have system-missing values, since *any* character is legal in a string variable.

Forecasthtg Error: The discrepancy between an observed value and its foricast, based on a specified model.

In SPSS: **To Create a New Time Series Variable**

From the menus choose: 1- Transform 2- Create Time Series...

-Select the time series function you want to use to transform the original variable(s).

- Select the variable(s) from which you want to create new time series variables. Only numeric variables can be used.

- **Optionally, you can:**

- Enter variable names to override the default new variable names.
- Change the function for a selected variable.

In SPSS: Default new variable names are the first six characters of the existing variable used to create it, followed by an underscore and a sequential number. For example, for the variable ECONOMETRIC, the new variable name would be ECONOM_I .

In SPSS: **To Define Missine Values for a Variable**

- Make the Data Editor the active window.
- If the Data view is displayed, double-click the variable name at the top of the column in the Data view or click the Variable View tab.
- Click the button in the Missing cell for the variable you want to define.
- Enter the values or range of values that represent missing data.

Missing values for string variables: All string values, including null or blank values, are considered valid values unless you explicitly define them as missing. To define null or blank values as missing for a string variable, enter a single space in one of the fields for Discrete missing values. You cannot define missing values for long string variables (string variables longer than eight characters).

In MINITAB: **Missing value symbol:** Minitab uses an asterisk (*) in (numeric columns) **and a blank in (alpha columns) [I] to represent missing values in a column. When you enter * as a value in the Data window or at the DATA> prompt in the Session window, you do not have to enclose it in quotation marks. However, you must enclose the * in quotation marks when it** is **part of the command line. Most commands exclude from aftalysis all rows with a missing value and display the number of excluded points. When an arithmetic command operates on a missing value, Minitab sets the result to *.**

Sample from Columns

Randomly samples rows from one or more columns. You can sample with replacement (the same row can be selected more than once), or without replacement (the same row is not selected more than once).

Dialog box items

Sample rows from column(s): Specify the number of rows to randomly select, and then select the column(s) you want the data to be sampled from.

Store samples in: Specify the column(s) where the sampled values are to be stored If **you sample from several columns at once, they must all have the same length; Minitab selects the same rows from each column.**

Sample with replacement: Check to sample with replacement. If left unchecked, Minitab samples without replacement, so the same row cannot be selected more than once.

To select a Random Sample from columns.

You can randomly sample rows from one or more columns.

I Choose Cale > Random Data > Sample from Columns.

2 In Sample, enter the number of rows you want to sample.

3 In the box following rows from column(s), enter equal-length Columns from which to sample.

4 In Store samples in, enter the columns in which you want to store the sample data. Click OK.

[I] **-Text data:** **Minitab handles numeric data, text data (formerly called aloha data), and date/time data.**

Rules for text data include:

I - Text data in a column may be up to 80 characters long.

2 -You may use any characters (letters, numbers, punctuation symbols, and blanks).

3 - No column can contain both text and numeric data:

Minitab treats numbers appearing in any text column as text characters.

Some Minitab commands do not handle text data. You can convert text to other types . of data.

(2) ESTIMATION METHODS for REPLACING MISSING VALUES

Suppose that the values of a time series are:

X1, X2, X3, ... X10,

Missing observations can be problematic in analysis, and some time series measures cannot be computed if there are missing values in the series. Replace Missing Values creates new time series variables from existing ones, replacing missing values with estimates computed using one of these methods.

(2-1) - Series mean.

Replacing missing values with the mean for the entire series.

All missing values are replaced by (K) where:

K = The mean of time series =E /n , 2, ..., n (I)

(2-2) - Mean of nearby points.

Replacing missing values with the mean of valid surrounding values.

The span of nearby points is the number of valid values above and below the missing value used to compute the mean.

Let X_i is missing value. Then X1 is replaced by (K) where:

1(FE PC1.2 + +X,,21 / **4** ...... **(2)**

(If the span of nearby points = 2), and 2 < i< n-2;

K FE PCi4+ XI-3 4X1-2+ XI-I+ X1+1 +X1+2 4 1 1+3 +44] /8 (3)

(If the span of nearby points = ***0),*** **and 4 < i < n-4;**

In general:

K =1.(Xi-+ Xi4g.1) + ... +X;.1+ + X;42+ + X1,,t1 / (2..) **...(4)**

(If the span of nearby points = g) **and g < i < n-g**

(2-3) - Median of nearby points.

Replacing missing values with the median of valid surrounding values.

The span of nearby points is the number of valid values above and below the missing value used to compute the median.

Let X1 be the missing value. Then X; is replaced by (K) where:

K = median of (X1.2, X1.1, (5)

(If the span of nearby points = **2), and 2 <i <** n-2;

K = median of Pl14, X1.3, XI-2, X1-1, X1+12 Xi+2, XI+3, X;+41 .... (6)

(If the span of nearby points = 4), and 4 <i < n-4;

In general:

K = median of [X1.5, Xim X1+2, .(7)

(If the span of nearby points = **g) and g < i < n-g**

(2-4) - Linear interpolation.

Replacing missing values usine a linear interpolation.
The last valid value before the missing value and the first valid value after the missing value are used for the interpolation. *If the first or last case in the series has a missing value, the missing value is not replaced.*

Let X; is missing value. Then Xis replaced by (K) where:

K = Xi., + - X(.11 / 2 ....(8)

2K = 2 X,.1 + X1+1- Xi4 X1-1 Xj+1 .... 9)

Then, K = +X.11/ 2 which I< i 12-1 ....(10)

But note that:

a - Linear interpolation = Mean of nearby points
(If the span of nearby points = 1, and there is no sequential values are missing).

b - If X. X., are missing values .Then Xi, Xi+, are replaced by (Ri, K1+1) where:

K;= 3`1+ X1.11/ 3 11)

= + (1/3)* IX02- Xi-il; ....(12)

(X.2- Xi.11 ....(13)

+ 1 3 - .... (14)

Ki+1 = Ki - Z2/ 3rriXin

In general:

If Xi, Xj+1, ... , X. are missing values.
Which: m = number of sequential missing values, and 1 < m; n-1
Then Xi, Xp2, ... , X. are replaced by uq, Ki+1,............ , IC,,,,)
Where: 2 .5 j < m -1

Then:

Ki = +(1 / (m+1)* Xi-il, ... (15)

Xj- i + (2 / (m+1)* Xi-11 ....(16)

= KJ +(1 / (m+1)* iXtp+I Xj-Il ....(17)

And:

K. =Xj., + (m / (m+I)* /X.+, — X01 ....(18)

+ (1 / (m+1)* (X.+1— ....(19)

(2-5) - Linear trend at point.

Replacing missing values with the linear trend for that point.
The existing series is regressed on an index variable scaled 1 to n.
Missing values are replaced with their predicted values.

The steps of Replacing Missing Values:

I - Suppose that the observations of a time series (missing and nonmissing values) are:
X1, X2, X3, ... X0-2, Xp-I, X.
And consider X is a dependent (response) variable

II — Let Y be an independent (predictor) variable, and Y' s values are: 1, 2, 3, ...

III — after excluding the rows of missing values from X, Y columns,

Determine the regression equation: X = a + b*Y

Let X, is missing value. Then Xi is replaced by (K ;) where:
K, = a + b*i
.... (20)

Missing Values in Functions: Functions and simple arithmetic expressions treat missing values in different ways.

In the expression: (varl+var2+var3)/3

The result is missing if a case has a missing value for at of the three variables.

In the expression: MEAN (varl, var2, var3)

The result is missing only if the case has missing values for all three variables.

For statistical functions, you can specify the minimum number of arguments that must have nonmissing values. To do so, type a period and the minimum number after the function name, as in
MEAN.2 (varl, var2, var3)

In SPSS: To Replace Missing Values for Time Series Variables
From the menus choose:

- **Transform**
- **Replace Missing Values...**
  - Select the estimation method you want to use to replace missing values.
  - Select the variable(s) for which you want to replace *missing* values.
- **Optionally, you can:**
  - Enter variable names to override the default new variable names.
  - Change the estimation method for a selected variable.

(3) TIME SERIES TRANSFORMATION FUNCTIONS

In this section we will illustrate the meaning of the system-missing value for the variable in many cases of transformations:

I - Difference. Non seasonal difference between successive values in the series. The order is the number of previous values used to calculate the difference. Because one observation is lost for each order of difference, system-missing values appear at the beginning of the series. For **example,** if the difference order is 2, the first two cases will have the system-missing value for the new variable.

2 - **Seasonal difference.** Difference between series values a constant span apart. The span is based on the currently defined periodicity. To compute seasonal differences, you must have defined date variables (Data menu, Define Dates) that include a periodic component (such as months of the year). The order is the number.of seasonal periods used to compute the difference. The number of cases with the system-missing value at the beginning of the series is equal to the periodicity multiplied by the order. For example, if the current periodicity is 12 and the order is 2, the first 24 Cases will have the system-missing value for the new variable.

3 - **Centered moving average.** Average of a span of series values surrounding and including the current value. The span is the number of series values used to compute the average. If the span is even, the moving average is computed by averaging each pair of uncentered means. The number of cases with the system-missing value at the beginning and at the end of the series for a span of n is equal to n2 for even span values and for odd span values. For example, if the span is 5, the number of cases with the system-missing value at the beginning and at the end of the series is 2.

4 - **Prior moving average.** Average of the span of *series* values preceding the current value. The span is the number of preceding series values used to compute the average. The number of cases with the system-missing value at the beginning of the series is equal to the span value.

5 - **Running median.** Median of a span of series values surrounding and including the current value. The span is the number of series values used to compute the median. If the span is even, the median is computed by averaging each pair of uncentered medians. The number of cases with the system-missing value at the beginning and at the end of the series for a span of n is equal to n/2 for even span values and for odd span values. For example, if the span is 5, the number of cases with the system-missing value at the beginning and at the end of the series is 2.

6 - **Cumulative sum.** Cumulative sum of series values up to and including the current value.

7 - **Lag.** Value of a previous case, based on the specified lag order. The order is the number of cases prior to the current case from which the value is obtained. The number of cases with the system-missing value at the beginning of the series is equal io the order value.

8 - Lead. Value of a subsequent case, based on the specified lead order. The order is the number of cases after the current case from which the value is obtained. The number of cases with the system-missing value at the end of the series is equal to the order value.

9 - Smoothing. New series values based on a compound data smoother. The smoother starts with a running median of 4, which is centered by a running median of 2. It then resmoothes these values by applying a running median of 5, a nt, ning median of 3, and harming • (running weighted averages). Resit& •1s are computed by subtracting the smoothed series from the original series. This whole process is then repeated on the computed residuals. Finally, the smoothed residuals are computed by subtracting the smoothed values obtained the first time through the process. This is sometimes referred to as T4253H.smoothing.

(4) MEASURES of ACCURACY

Accuracy refers to the closeness of the measurements to the "actual" or "real" value of the physical quantity. Therefore, an "accurate" estimate has small bias.

Minitab computes three measures of accuracy of the fitted model: **MAPE, MAD, and MSD** for each of the simple forecasting and smoothing methods. For all three measures, the smaller the value, the better the fit of the model. Use these statistics to compare the fits of the different methods.

(I) - Mean Absolute Percentage Error (MAPE)

MAPE measures the accuracy of fitted time series values. It expresses accuracy as a percentage.

MAPE = |(Yr i',)/ Y, | + n * 100 , ... (21)

Where: (Y) equals the actual value; (/7) equals the forecast value; **(n)** equals the number of forecasts; t = 1, **2,, n; and Y, # 0**

- Mean Absolute Deviation (MAD)

MAD measures the accuracy of fitted time series values. It expresses accuracy in the same units as the data, which helps conceptualize the amount of error.

MAD=E1(Yt-t)i +n .... (22)

- Mean Squared Deviation (MSD) in Time Series

MSD is very similar to MSE, mean squared error, a commonly-used measure of accuracy of fitted time series values. Because MM) is always computed using the same denominator, n, regardless of the model, you can compare MSD values across models. Because MSE" are computed with different degrees of freedom for different models, you cannot always compare MSE values across models.

$$\text{MSD} = (\text{V},\text{- to}^{2}\ \text{n} \qquad \text{............}(23)$$

(5) – NUMERICAL STUDY

Methods dealing with analysis of data with missing values (Incomplete Data) can be classified into:
- Analysis of complete cases, including weighting adjustments,
- Imputation.methods, and extensions to multiple imputation, and
- Methods that analyze the incomplete data directly without requiring a rectangular data set, such as maximum likelihood and Bayesian methods.

In this section:

- We conduct simulations designed to investigate what the effect on the forecasting accuracy of the time series model, when using various Estimation Methods for Replacing Missing Values, various probability distributions, and various sample sizes.
- The simulations were carried out using two statistical packages (MINITAB and SPSS) .The random data generator used in the simulations is that from the MINITAB calc Toolbox, and the Estimation Methods for Replacing Missing Values used in the numerical study is that from the SPSS Transform Toolbox.
- We choose three continuous Distributions (2 symmetric distributions and one assymmetric distribution), and one Discrete Distribution for our simulation study.
- Our symmetric distributions are the standard normal distribution N (0, 1), and the Student t — distribution with 5 degrees of freedom [T-5]. These represent moderate and very heavy tailed distributions.
- For the assymmetric distribution we use the log-normal distribution, with is = 0, o =1, denoted Log-N (0, 1). (A variable X has a lognormal distribution if log (X) has a normal distribution), and

For the Discrete Distribution we use the Binomial Distribution. (The probability of success = 0.5),

- For a range of values for n (50, 100, 500, *and 1000),* we use the four distributions and calculate the results of the various Estimation Methods for Replacing Missing Values discussed in the section-2, and the various Measures of Accuracy discussed in the previous section.
- For the Mean of nearby points and Median of nearby points *as* Estimation Methods for Replacing Missing Values we choose the span of nearby points = **4.**
- For all cases, always we consider that the number of the missing values equals to twenty percent of sample size, then (m) = (20%*n)

<u>The steps of numerical study:</u>

1 - Using MINITAB, to generate the random data according to the probability distribution and the sample size (n),.

2- Select a Random Sample (m) from the column of the sample size in the step one, and do it *as* missing values, where (m) = (20%*n).

3- Using SPSS , to estimate the Missing Values by using the previous Estimation Methods.

4- Compute the measures of accuracy using:
a - MINITAB Toolbox —0 cafe-0 calculator, or
b - SPSS Toolbox —. Transform—. compute

5 — Repeat the steps 2, 3, and 4 for another random sample from the same column (the same probability distribution and the same sample size).

6 - Compute the average of the computed measures of accuracy for all selected samples. *In this paper, we select 100* Random Samples in each case.

7 - Repeat the steps 2, 3, **4, 5, and** 6 for all cases.

- **Our simulation results are reported in the following Tables: (1, 2, 3, and 4).**

If N = population size, and n = sample size. Then we can select number ofsimple random samples (without replacement) = $^{N}C_{,,}$ *or (N) Combination (n).*

Table (1): n = 50

Estimation Methods for Re lacin Missin: Values

Probability distribution	Measures of Accuracy	Series mean	Mean of nearby points(4)	Median of nearby oints(4	Linear interpolation	Linear trend at point
N0, 1)	MAPE	0.9902	2.3228	2.5416	2.4054	.9886
	MAD	0.6782	0.8014	0.7723	1.1395	(0.6781
	MSD	0.3772	0.3999	0.2900	1.9232	0.3753
T-5	MAPE	0.3015	0.1108	0.7727	0.9422	0.1752
	MAD	1.4889	1.5565	1.9242	1.9950	1.4473
	MSD	.9213	5.2079	6.6786	4.7922	42257
Log-N (0, 1)	MAPE	9.5667	10.4932	7.5710	15.0341	9.6915
	MAD	1.5616	1.5523	1.3906	.3766	1.5615
	MSD	1.6258	1.0021	1.0615	1.0982	1.6237
Binomial	MAPE	0.0987	0.1899	0.1289	0.3219	0.1087
	MAD	3.4297	5.0143	4.9877	52995	3.2934
	MSD	6.9329	11.2396	9.1002	5.2062	8.008

Table (2): n = 100

Estimation Methods for Re lacin Missin Values

Probability distribution	Measures of Accuracy	Series mean	Mean of nearby points(4)	Median of nearby oints(4)	Linear interpolation	Linear trend at point
N (0, I	MAPE	0.4540	0.4934	0.4331	0.3912	0.4347
	MAD	1.0057	1.0822	1.0657	1.2133	(0.9988
	MSD	1.3421	1.3821	1.0942	(0.9231	1.3286
T-5	MAPE	0.1051	awn	0.2711	0.4920	0.0720
	MAD	1.3568	1.4931	1.4281	1.5070	1.3720
	MSD	3.1731	4.7092	7.6423	4.2971	3.5793
Log-N (0, 1)	MAPE	2.0350	2.0217	1.2838	2.5772	2.0473
	MAD	(3.6306	3.6579	3.7968	3.6554	3.6885
	MSD	1.5652	1.6117	2.3743	1.2675	1.7640
Binomial	MAPE	0.0782	0.0959	0.0939	0.1196	0.0778
	MAD	3.9784	4.8438	4.7000	5.9500	3.9347
	MSD	8.0391	13.0008	7.2000	5.2200	8.8337

Table (3): n = 500

Probability distribution	Measures of Accuracy	Estimation Methods for Re • lacing Missin• Values				
		Series mean	Mean of nearby points(4)	Median of nearby points(4)	Linear interpolation	Linear trend at point
	MAPE	0.8264	0.5845	0.6714	0.3255	***(0.3153)***
N (0, 1)	**MAD**	0.9246	0.9881	1.0223	1.1242	***(0.9145)***
	MSD	1.1322	1.1841	0.9846	***(0.6244)***	1.3888
	MAPE	0.1396	0.4179	0.2195	***(0.0106)***	0.1442
T-5	**MAD**	***(1.1160)***	1.2039	1.1650	1.21714	1.1176
	MSD	2.2299	2.4764	2.7124	***(1.39526)***	2.1924
	MAPE	1.8195	2.0320	***(1.5427)***	2.2643	1.8372
Log-N (0, 1)	**MAD**	***(2.5730)***	2.750	2.735	2.877	2.595
	MSD	***(1.7398)***	1.7693	2.7794	1.7322	1.7668
	MAPE	0.0902	***(0.0865)***	0.0916	0.1200	0.0904
Binomial	**MAD**	4.6660	***(4.4545)***	4.6364	6.2727	4.6856
	MSD	25.0020	24.7500	***(20.4545)***	105.0909	26.4712

Table (4): n = 1000

Probability distribution	Measures of Accurac	Estimation Methods for Re • lamg Missing Values				
		Series mean	Mean of nearby •ints 4	Median c4'nearby oints(4)	Linear interpolation	Linear trend at int
N (0, 1)		0.9298	0.3648	0 3927	0.2340	***(0.0536)***
	MAD	***(0.9324)***	.0.9950	1.0154	1.0527	0.9450
	MSD	0.2947	0.2968	0.5864	***(0.1182)***	0.2873
						IMO
	MAPE	0.1113	0.3141	0.2211	***(0.0513)***	0.1004
T-5	MAD	1.1086	**1.3334**	1.2036	**1.1784**	***(1.0917)***
	MSD	2.0099	**2.7158**	2.9257	***(14329)***	1.9919
	MAPE	2.1819	273035	***(1.7024)***	2.6892	1.9283
Log-N (0, 0	MAD	***(26973)***	2.7995	2.9542	2.9998	2.8195
MINE	MSD	1.9812	2.0195	2.0019	***(1.8998)***	1.9976
	MAPE	0.0926	[illegible]	0.1209	0.2151	0.0897
Binomial	MAD	4.3397	refintEES		5.9958	4.3193
111.11111	MSD	28.0103	27.0098	***17.2101***	[illegible]	28.9983

- m the above tables each cell has the estimated values of PE, MAD, and MSD for each Estimation Method
- any case (row), the best Estimation is ***Bold, Italic· and in brackets.***

6 — *CONCLUSION*

Our simulation results demonstrate that:

I — For the symmetric distributions:

- With normal distribution, we can recommend the Linear trend at point or the linear interpolation method in most cases, especially when the sample size is large (n > 50).
- With t distribution (heavy tailed distributions):

  1 — For n < 500, we can recommend the Series mean method or the Mean of nearby points (the span of nearby points = 4).

  2 — For n > 500, we can recommend the linear interpolation method.

II - For the assymmetric distributions (log-normal distribution):

1 — For n < 500, we can recommend the Series mean method or the Mean of nearby points (the span of nearby points = 4).

2 — For n > 500, we can recommend the linear interpolation or the Median of nearby points method (the span of nearby points = 4).

- In general, for the continuous Distributions, we can recommend the linear interpolation method if the sample size is large. Also we can recommend the Series mean method if the sample size < 500.

III — For the Discrete Distributions (Binomial distribution):

1 — For n < 500, we can recommend the linear trend at point or the Linear interpolation method.

2 — For n > 500, we can recommend the Mean of nearby points or the Median of nearby points (the span of nearby points = 4).

Finally, we recommend in the future studies that changing:

1 - The sample size (n).
2 - The span of nearby points (any even or odd number).
3 — 7 he probability distributions.
4 — The number of random samples, which *we select* in each *case (>100).*
5 - The number of missing values in the selected sample (m # 20%•n).

---

7 — REFERENCES

1 - Rubin D., *"Multiple Imputation for Nonresponse in Surveys ",* New York, Wiley, 1987.

2 - Schafer J., *"Statistical Analysis with Missing Data",* London, New Chapman and Hall, 1997.

3 — Manual of SPSS - v.11 (statistical package software).

4 - Manual of MINITAB - v.14 (statistical package software).

5 -Internet Websites, which are related with this paper subject.

***