International Journal of Economics and Finance; Vol. 10, No. 10; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Deriving Correlation Matrices for Missing Financial Time-Series Data

Schalk Burger[1], Searle Silverman[2] & Gary van Vuuren[1]

[1] Centre for BMI, North-West University, Potchefstroom Campus, South Africa

[2] RMB, Johannesburg, South Africa (formerly Deloitte)

Correspondence: Schalk Burger, Centre for BMI, North-West University, Potchefstroom Campus, South Africa, Tel: 27-071-488-4973. E-mail: sburger29@gmail.com

Received: May 16, 2018 Accepted: September 23, 2018 Online Published: September 28, 2018

doi:10.5539/ijef.v10n10p105 URL: https://doi.org/10.5539/ijef.v10n10p105

Abstract

The problem of missing data is prevalent in financial time series, particularly data such as foreign exchange rates and interest rate indices. Reasons for missing data include the clo-sure of financial markets over weekends and holidays and that sometimes, index data do not change between consecutive dates, resulting in stale data (also considered as missing data). Most statistical software packages function best when applied to complete da-tasets. Listwise deletion – a commonly-used approach to deal with missing data, is straightforward to use and implement, but it can exclude large portions of the original dataset (Allison, 2002). Where data are randomly missing or if the deleted data are insignificant (measured by statistical power), listwise deletion may add value. Techniques to handle missing data were suggested and implemented. These techniques were assessed to ascertain which provided the most accurate reconstructed datasets compared with complete dataset.

Keywords: missing data, correlation, Stineman, interpolation

1. Introduction

1.1 Background

The problem of missing data is widespread and poses a problem in financial time series data such as foreign exchange (FX) and interest rates. Data may be absent from financial time series for various reasons: financial markets close on weekends and holidays (with the latter being different in different countries) and indices sometimes do not change value over certain periods (called "stale data") and considered as "missing data" for this paper. Missing data pose a serious problem for statistical analysis which requires complete datasets. For example, the Pearson correlation requires the same number of pairwise data for its computation: errors result if this constraint is not satisfied through missing data.

Several techniques exist to handle missing or stale data. Listwise deletion (Note 1) (Complete Case Analysis), a commonly-used method is easy to use and implement. Although the simplicity provides an advantage, some significant disadvantages exist such as the potential exclusion of a large proportion of the original data (Allison, 2002). It can be a valuable technique when data are missing randomly (Note 2&3), or if the cases deleted do not result is significant differences between calculated values.

The scope of this paper is to investigate, evaluate and test different techniques to deal with missing data and then apply these techniques to financial time series data to construct correlation matrices. The statistical software R Studio (Note 4) was used to implement, test, and evaluate the different techniques used to handle missing data.

The remainder of the paper is structured as follows. A review of the available literature is provided in Section 1, followed by a description of the data used and the methodology employed in Section 2. The results are documented in Section 3 and Section 4 concludes.

1.2 Literature Review

Missing data patterns: A missing data pattern refers to the configuration of observed and missing values in a dataset (Enders, 2010). Little and Rubin (2002) define missing data patterns as those which describe the observed and unobserved values in a data matrix. Missing data pattern analysis therefore focuses on the patterns themselves, not the reason why the data are missing (called missing data mechanisms). These represent two distinct issues.

Honaker et al. (2011) introduce a practical aspect of missing data patterns using R software (Honaker et al., 2011), which provides a visual tool to represent missing data pattern in time series data. Six possible missing data patterns are described below.

Univariate Pattern: one where the missingness is isolated to a single variable (Little & Rubin, 2002). This pattern is rare but arises in experimental studies or designed experiments (Enders, 2010).

Unit Nonresponse Pattern: occur in survey research and can occur when respondents refuse to answer several questions (Enders, 2010). For example, question Y_3 and Y_4 have missing values if the columns in Figure 1 are questions or a series of questions.

Monotone Pattern: frequently accompanies a longitudinal study and are often encountered in healthcare. They are sometimes referred to as *attrition in longitudinal studies* (Little & Rubin, 2002), in which subjects are continuously monitored over long periods according (Trueman, 2016). Schafer (1998) argues that these patterns reduce the mathematical complexity of maximum likelihood and multiple imputation and can eliminate the need for iterative estimation algorithms. Schafer (1998) states that the reason for missing variables is subjects dropping out of the study in an early stage or subjects being unavailable for a period during the study.

General Pattern: the most common missing data pattern (Enders, 2010). The values may seem completely random at first, but values can still be missing according to a specific pattern.

Planned Missing Data Patterns: can be used for collecting many questionnaire items while simultaneously reducing burden on respondents (Enders, 2010).

Latent Variable Pattern: reflects variables that are completely absent (Little & Rubin, 2002), and is calculated with missing data theory. These patterns are unique to latent variable analyses such as structural equation models (Enders, 2010).

The white shaded areas in Figure 1 indicate available data observations while the grey shaded areas reflect missing data for the missing data patterns described above.

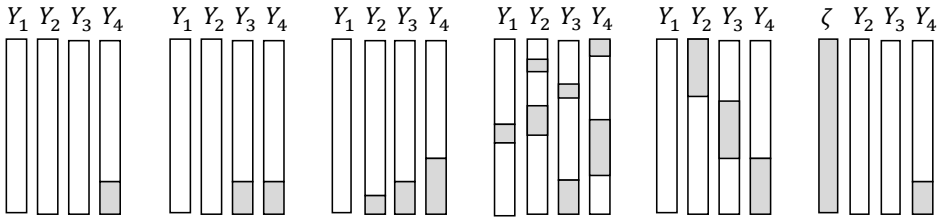


Figure 1. Missing data patterns. From left to right: univariate, unit non-response, monotone, general, planned missing data and latent variable patters (Enders, 2010)

1.3 Classification of Missing Data

Rubin (1976) first introduced a classification system for missing data, by treating the missing data indicators as random variables and assigning a distribution to them. Little and Rubin (2002) continued the idea of mechanisms that lead to missing data. They reviewed the theory first introduced by Rubin (1976), using different notation and terminology compared to those of the original paper. Enders (2010) elaborated on the classifications systems and the fact that missing data mechanisms do not necessarily provide causal explanations for the missing data, but that they do represent the generic mathematical relationships between the data and missingness. Missing data mechanisms can be described as possible relationships between measured variables and the probability of missing data (Enders, 2010).

The definitions that follow are derived from Little and Rubin (2002) and Enders (2010).

The missing indicator is denoted as R and the missing data mechanisms describe the different relationships between R and the data. Y_{obs} are the observed parts of the data and Y_{mis} are the missing parts. The parameter (or set of parameters), ϕ, describes the relationship between R and the data. The missing data mechanisms are defined formally below.

Missing Completely at Random (MCAR) requires that missingness is completely unrelated to the data, sometimes referred to as haphazard missingness (Enders, 2010). Little and Rubin (2002) emphasise that this assumption does not mean that the pattern itself is random, but that the missingness does not depend on the data

values, Y, that includes Y_{mis} and Y_{obs}. The probability distribution of MCAR is:

$$p(R|\phi) \quad (1)$$

Missing at Random (MAR) is defined as the missingness that depends on Y_{obs} but not on Y_{mis} (Graham, 2012). Occasionally, MAR is defined as a missingness mechanism in which there is no relationship between the propensity for missing variables on Y and the values of Y are partialling (the process by which a single variable is assigned a fixed value to identify correlations between the other variables) out other variables (Enders, 2010). MAR does not mean that the data are missing randomly, but that a systematic relationship between one or more measured variables and the probability of missing data exist (Enders, 2010). In simple terms this means that the missingness was not caused by a completely random process (Graham, 2012). The probability distribution of MAR is:

$$p(R|Y_{obs}, \phi) \quad (2)$$

Missing Not at Random (MNAR) is also known as Not Missing at Random (NMAR); these terms can be used interchangeably (Little & Rubin, 2002; Enders, 2010). It is sometimes defined as the type of missingness in which the cause of missingness is correlated with Y (Graham, 2012). The variable Y can sometimes be missingness or contain missingness depending on the way it is measured. Enders (2010) describes MNAR as a missingness mechanism were the probability of missing data on a variable Y is related to the values of Y itself. The probability distribution of MNAR, defined below, is useful because it contains all the information about the missingness:

$$p(R|Y_{obs}, Y_{mis}, \phi) \quad (3)$$

where ϕ is a parameter that describes the relationship between R and Y (Y_{mis} and Y_{obs}).

1.4 Imputation of Missing Data Approaches

Little and Rubin (2002) provide information about statistical analysis from missing data experts. A comprehensive overview of data editing and imputation may be found in De Waal, Pannekoek, and Scholtus (2011). Although not necessarily an exploration of the niche financial time series field, it does provide a comprehensive overview of the topic. Enders' (2010) covers applied missing data analysis and translates state-of-the art technical missing data literature into an accessible, reliable reference.

The approach used in this paper to address the problem of missing data is "multiple imputation", a common approach employed in several fields. Statistical models are used to extract relevant information from observed data sets and to use this to impute several values for the missing values. Several "complete" data sets are thus produced with *observed* values the same in each, but the imputed values differ depending on the approach used. The appeal of the process is that, after imputation, statistical methods (which would have been employed had there been no missing values) can be applied to each of the completed data sets. Simple procedures can then be applied to combine results.

The most common statistical analysis methods require data sets with no gaps. Real, empirical data, however, are replete with gaps and scattered missingness throughout. A possible circumvention of the problem is to employ listwise deletion, a process by which any record/observation or case is excluded from an analysis if any single observation in the record is missing (Graham, 2012). Listwise deletion discards all the (often substantial) information which exists in the partially-observed observations and which encodes the variable's relationships. Better solutions than listwise deletion involve plugging the gaps with statistical estimates, but since missing data cannot be replaced with the "true" values of the missing data (in which case there would be *no* missing data), missing data must be imputed. Imputed values, however, cause statistical analysis software to exaggerate confidence in the output (by biasing standard errors and confidence intervals), more so than is justified because there are fewer data than were empirically observed.

The missing data theory that preceded this section lays the foundation for understanding the key concepts necessary to understand the techniques to treat missing values. In this paper, daily returns will be used for statistical analysis. Closing prices were used for the analysis. According to Taylor (2008) daily returns are more convenient to analyse changes in prices with than direct statistical analysis of financial prices, because consecutive prices are highly correlated, and the variances of the prices increase with time. This increase in prices is due to non-stationarity and the daily returns makes the process stationary.

The techniques to handle missing values that are listed below have been identified using a pragmatic approach. Hastie et al. (2009) lists three ways to handle missing variables, assuming they are MCAR:

- discard observations with any missing values,

- rely on learning algorithm to treat missing values, and
- impute all missing values before using them in any calculations.

Next the techniques that can be used to handle missing values will be described. Each technique will be discussed briefly in terms of the mechanics behind the technique, the advantages, disadvantages and implementability of each technique.

According to Honaker et al. (2011), these approaches can lead to serious biases and covariances. *Ad hoc* techniques to treat missing data, for example listwise deletion, tend to discard information about variables to make the estimation problem more tractable (Schafer, 1997). This could lead to a loss in statistical power and listwise deletion may be biased if the missingness mechanism is MAR and not MCAR.

The *ad hoc* method of mean imputation may preserve the observed sample mean, but it distorts the covariance structure, biasing estimated variances and covariances toward zero. Imputing variables from a regression model, inflates the correlations, biasing them away from zero. When the missingness mechanism is complex, the derivation of imputation scheme that preserves the important aspects of a joint distribution can prove to be very difficult (Schafer, 1997).

Ad hoc approaches are, therefore, not necessarily a good way of treating missing values as these approaches are outdated.

Listwise deletion means that the variables with missing data on any of the variables used for the statistical analysis, will be excluded from the analysis. The advantages of this technique include that it is easy to implement and it is the default setting of most of the statistical software programs. Thus, the implementation of these techniques is straightforward. The disadvantages are however that this technique can greatly reduce the sample size that is used for statistical analysis. There is also the risk of potential bias in the mean as well as the variance of the statistical analysis on the data.

Mean imputation a simple method which is easy to understand and implement in any statistical software program. This method replaces all the missing values with the mean of the observed values of each variable, thus the mean of the variables observed is not biased and remains unchanged with this method. The disadvantage of this method is that the variance of the data is reduced.

Multiple Imputation (MI) was designed by Rubin (1976) to be practical without neglecting statistical elements. MI reduces the bias and increase efficiency compared to *ad hoc* approaches (Honaker *et al*., 2012). MI is very flexible and can be used with any kind of model or data, but it is a difficult process to implement due to the extremely technical nature of the process and subsequent algorithms. Another drawback of MI is that a different result is obtained every time it is used.

Maximum Likelihood (ML) is preferred to Multiple Imputation (MI) because ML is consistent and asymptotically efficient under the MAR assumption (Allison, 2012). If the missingness mechanism can be describe by MAR then the standard errors are unbiased. The ML method can be used incorporating the Expectation Maximisation (EM) algorithm. Dempster et al. (1977) first introduced the EM algorithm. The EM Algorithm follows a two-step approach to get ML estimates from the missing data:

- Expectation (E): Calculate the expected value of the log-likelihood for observed data, based on current parameter estimates.
- Maximisation (M): Maximise the expected likelihood to obtain new parameter estimates.

The above process is repeated until convergence is obtained.

Missing data patterns and missing data mechanisms should be considered when dealing with missing data. When using a technique to handle the missing data the missing data pattern must be checked to see which pattern the missingness belongs to. This can provide valuable insight into the data in cases when data is not missing at random or when a specific pattern is evident in the missing data. Missingness mechanisms can influence the techniques used to handle the missing data, thus the missingness mechanisms must be tested if it is not known. When known, the choice of techniques to handle missing data can be made with more certainty.

2. Method

2.1 Data

Two datasets were used for the implementation of the different techniques to handle missing data, one from Quandl (only FX data) and one from Bloomberg (both FX and interest rate data). Each dataset differs in terms of the number of observations, the number of missing data and variables. The two datasets – which comprised daily

data for 10.5 years (Jan 06 – Jun 17) for the Bloomberg dataset and daily data for six years (Jan 2011 – Dec 2016) for the Quandl dataset – were used for implementing and comparing the different techniques to handle missing data. These indices comprise 2 947 incomplete observations because of missing and stale data. The London Interbank Offered Rate (LIBOR) and Johannesburg Interbank Average Rate (JIBAR) are indexed rates that are used as benchmarks for determining interest rates. The data appear in Table 1 below.

Table 1. Data

Rate	Source	Details
Interest	Bloomberg	3-month LIBOR Switzerland (Swiss Franc LIBOR), 6-month Euribor (Euro LIBOR), 3-month LIBOR United Kingdom (British Pound LIBOR), 6-month LIBOR Japan (Japanese Yen LIBOR), 3-month LIBOR United States (US Dollar LIBOR), 3-month JIBAR, 3-month LIBOR Australia (Australian Dollar LIBOR).
FX	Quandl	AUD:USD, CHF:USD, EUR:USD, GBP:USD, JPY:USD, ZAR:USD AUD:USD, CAD:USD, CHF:USD, CNY:USD, EUR:USD, GBP:USD, HKD:USD, INR:USD, JPY:USD, MXN:USD

The most liquid interest indices (6-month EURIBOR, 3-month UK LIBOR and 3-month US LIBOR US) were used for testing and comparing the different techniques to handle missing data for both FX rates and interest rate indices.

Figure 2 indicates the extent of missing data with black vertical lines for two interest rates. The remainder of the data look similar, whether for interest or FX rates.

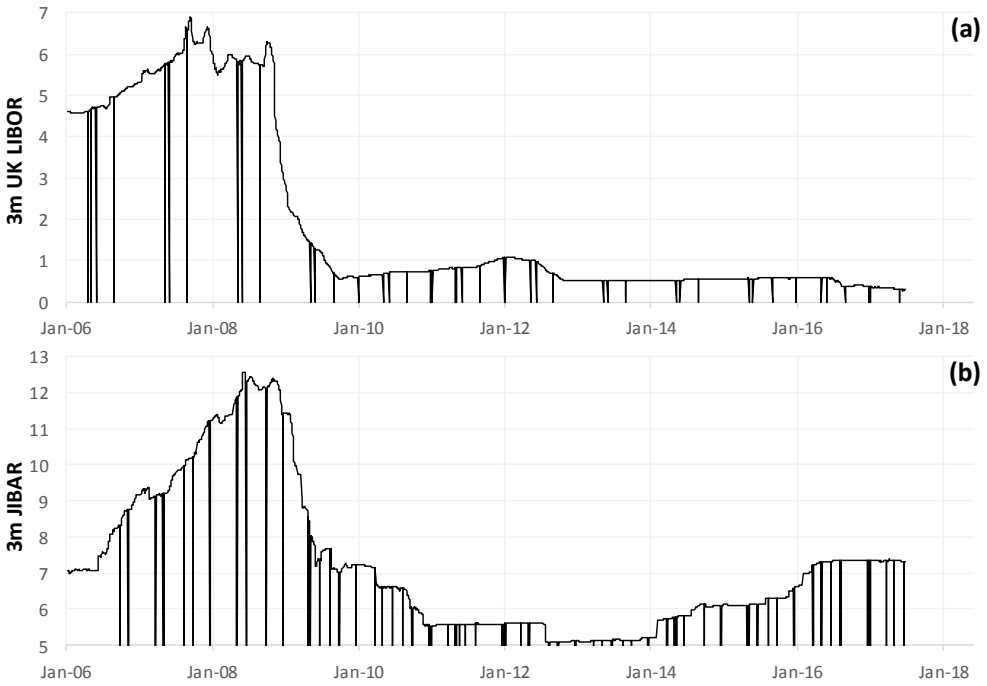


Figure 2. (a) UK 3m LIBOR and (b) 3m JIBAR interest rates, showing missing data as vertical lines

Table 2 shows the extent of missing data as a percentage of total missing data. Note that interest rates have more missing data than FX rates. Continuously-compounded daily returns were used for the calculations of the correlation matrix. These returns were used as opposed to simple returns, because of the additive property of the continuously compounded daily returns. The model was constructed in R Studio.

Table 2. Missing or stale data as a percentage of total sample data. The percentages for the same rates differ because of the different sample times

	Quandl		**Bloomberg**	
	FX rates		Interest rates	
AUD:USD	8.09%	0.78%	AUD	33.50%
CAD:USD	3.29%			
CHF:USD	3.33%	0.85%	CHF	47.00%
CNY:USD	3.79%			
EUR:USD	7.81%	0.54%	EUR	16.90%
GBP:USD	8.54%	0.37%	GBP	25.70%
HKD:USD	3.61%			
INR:USD	3.47%			
JPY:USD	3.33%	0.71%	JPY	44.80%
MXN:USD	3.47%			
ZAR:USD		0.10%	ZAR	72.70%
USD:USD			USD	23.30%

2.2 Methodology

The principle aim of this paper is to investigate, evaluate and test different techniques to deal with missing data and then apply these techniques to financial time series data with the ultimate aim of constructing correlation matrices (which are used for risk management, asset allocation, portfolio performance and regulatory capital calculations).

The correlation technique used in this paper is the Pearson correlation and correlations (Note 5) are calculated between FX rates and interest rate indices. The Pearson correlation coefficient is defined in Rice (2006) as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{4}$$

Where x_i is the ith observation of variable x, y_i is the ith observation of variable y, $\overline{x}$ is the sample mean of variable x, and $\overline{y}$ is the sample mean of variable y.

Pearson correlations are used to calculate market parameters used in Monte-Carlo (MC) simulations that simulate market scenarios. Market scenarios are used for calculating the potential future exposure (PFE) for counterparty risk and default risk. These risks can be used in fair value measurements as defined by IFRS (International Financial Reporting Standards)13 or regulatory capital as stipulated by the Basel Committee. To calculate the aforementioned, correlation matrices are needed. This paper focuses on the correlation matrices and the construction thereof with infrequently observable time series data, i.e. time series data with missing data. The scope does not include the calculation of market scenarios and subsequent steps to calculate the PFE for counterparty and default risk.

2.2.1 FX Rates

FX rates (Note 6) belong to the most efficient and liquid segments of financial markets (Tichy, 2006). Because of this property, FX rates do not usually have problems with stale data. Brownian motion is a continuous time version of a random walk with steps being normally distributed random variables (Rice, 2007). FX rates are assumed to follow a geometric Brownian motion (GBM). If Y_t is an exchange rate between two currencies at time t, Y_t behaves like a geometric Brownian motion, i.e. it follows a stochastic differential equation (SDE) of the form

$$dY_t = \mu Y_t dt + \sigma Y_t dW_t \tag{5}$$

Where Y_t is an FX rate, and W_t is a Wiener process.

2.2.2 Interest Rates

Interest rate indices (Note 7) are unsecured short-term borrowing rates between banks (Hull, 2012). JIBAR is an example of this. Interest rate indices are modelled using the Hull-White one factor model. The Hull-White one factor model, generalised by Vasicek with a time dependent parameter, is

$$dr = [\theta_t - a_t r]dt + \sigma_t \, dV_t \tag{6}$$

θ_t is a deterministic function of time, a is the mean reversion speed, σ is the volatility and V_t is a standard

Brownian motion under the risk neutral measure.

This paper employs both FX and interest rate indices for testing different techniques to handle missing data. The aim is to test the techniques that will not change the underlying distribution of the data with regards to the correlation matrix of the returns. In other words, the technique used to handle missing data with the correlation matrix "nearest" to that correlation matrix calculated with complete data is considered the "best" technique.

The correlation matrices constructed from the techniques to handle missing data were compared with the matrix from the *complete* dataset, by taking the absolute difference between the correlation matrices constructed by using the techniques and the complete case matrix. To illustrate this, a simplified example is used. Assume a correlation matrix is calculated from a complete dataset. Manually-created missing data are created and added to the dataset (there are various ways of doing this: imputation techniques and listwise deletion). This correlation matrix is then reduced to an upper triangle, as shown in Table 3.

Table 3. Example correlation matrix from complete dataset (and upper triangle)

	A	B	C			A	B	C
A	1	0.9	-0.5		**A**		0.9	-0.5
B	0.9	1	0.2	→	**B**			0.2
C	−0.5	0.2	1		**C**			

After the data have been treated with a technique to handle the missing data, a second correlation matrix is calculated (Table 4).

Table 4. Example correlation matrix from reconstructed data (and upper triangle)

	A	B	C			A	B	C
	1	0.7	-0.3		**A**		0.7	-0.3
B	0.7	1	-0.1	→	**B**			-0.1
C	−0.3	−0.1	1		**C**			

The absolute difference between the two correlation matrices is then calculated, creating in an upper triangle difference matrix. This matrix is an indication of "how far" away the two correlation matrices are from each other – shown in Table 5.

Table 5. Difference matrix between Tables 3 and Table 4

	A	B	C		A	B	C
A		|0.9 − 0.7|	|−0.5−(−0.3)|	**A**		0.2	0.2
B			|0.2 − (−0.1)| →	**B**			0.3
C				**C**			

Assume another imputation technique was used to treat the manually-created missing data. After following the approach outlined above, the difference matrix as shown in Table 6 was produced.

Table 6. Difference matrix from another data-imputation process

	A	B	C
A		0.1	0.3
B			0.2
C			

The differences in the correlation matrices (in Tables 5 and 6) are now summed to obtain a single value. From Table 5, this value is $0.2 + 0.2 + 0.3 = 0.7$. From Table 6 this value is $0.1 + 0.3 + 0.2 = 0.6$. The difference matrix with the lowest value is then the "best" difference matrix. Thus, to compare different techniques to handle missing data the single values (summed differences) were calculated and the lowest values identified. The lowest

values indicated which techniques are better than the other to handle the missing data. The aim of this paper is to ascertain which technique results in the best "difference matrix value" to handle missing data used for the calculation of correlation matrices.

These models are then compared with each other using evaluations to assess imputation accuracy. The root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to compare the different techniques to each other. All the above-mentioned comparison criteria are popular in the literature and may be used to compare and evaluate different techniques.

The RMSE is

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}, \tag{7}$$

Where $\hat{y}_t$ is the predicted value, y_t is the observed value, and n is the number of observations.

The MAE is

$$MAE = \frac{|\sum_{t=1}^{n}(\hat{y}_t - y_t)|}{n} \tag{8}$$

The MAPE is

$$MAPE = 100\frac{\left|\sum_{t=1}^{n}\left(\frac{(\hat{y}_t - y_t)}{y_t}\right)\right|}{n} \tag{9}$$

From these additional information criteria, the best techniques overall were chosen. To remove any bias, the tests were run over 1 000 simulations where missing data were generated, and the techniques used to handle the missing data were applied.

A schematic of the process thus followed is shown in Figure 3.

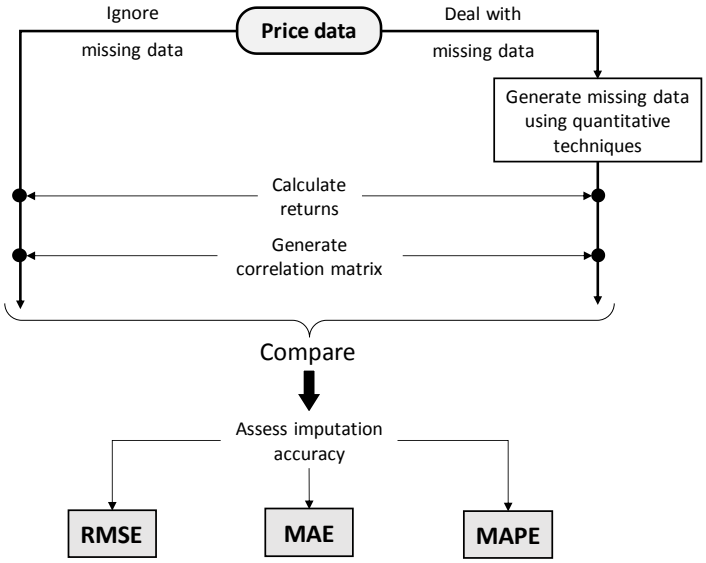


Figure 3. Simplified approach of model development and validation process

2.2.3 Imputation Techniques

Different imputation techniques were used to impute missing values. All methods were implemented, tested and compared with each other as well as with listwise deletion to ascertain which method imputed the best missing values used to construct the correlation matrices. A short description of each imputation technique follows:

Linear interpolation – the gap (missing data) between two points is replaced using a linear polynomial between the two known points.

Spline interpolation – a piecewise cubic (or Hermite) polynomial is fitted between the two known points to replace the missing data in between (second derivatives $= 0$ at endpoints) (Wagon, 2010).

Stineman interpolation – solves the non-monotonic problem of linear and spline interpolation. Each data point

has a slope obtained by fitting a circle to that point and its neighbours and using the slope of the tangent with the circle. Given the two points with assigned slopes a smooth function is obtained that connects the points and respects the slope (Wagon, 2010). The Stineman method is not as smooth as the linear and spline method (Stineman, 1980).

Mode imputation - replaces the missing data with the mode of the observed data.

Random Imputation – replaces missing data with a randomly drawn value from the observed data.

Mean Imputation - (also arithmetic mean or unconditional mean imputation) replaces missing data with the mean of the observed data. Mean imputation creates a complete dataset, but severely distorts the resulting parameter estimates, even if the data are MCAR (Enders, 2010). It also reduces the variability of the dataset and therefore will also affect the standard deviation and variance.

Median imputation – missing data are replaced by the median of the observed data.

Last observation carried forward - missing values are replaced with the previous observed value.

Next observation carried backward – missing values are replaced with the following observed value.

Moving average imputation – the moving average of the observed data replaces missing values.

Exponential Weighted Moving Average – missing data are replaced with the exponential weighted moving average of the observed data with a specified weighting parameter.

Linear Weighted Moving Average Imputation – missing data are replaced with the linear weighted moving average of the observed data.

***k* nearest neighbour imputation** – a non-parametric learning algorithm used by Google to autocomplete Google searches. k nearest neighbour imputation does not provide an exact match, but is a scenario driven method in which different scenarios are compared with each other. The difference between scenarios is a certain uniqueness and these scenarios are the neighbours. New scenarios are compared with each scenario in the model and matched according to the closest-neighbour to the case (Waqas et al., 2016). k in the name is the amount of neighbours each case comprises.

Random forest imputation – uses random forests to impute the missing data by running forests for many iterations.

Kalman imputation - uses Kalman smoothing on structural time series models or state space representation of an ARIMA model.

Listwise deletion – a process by which any record/observation or case is excluded from an analysis if any single observation in the record is missing (Graham, 2012). Listwise deletion discards all the (often substantial) information which exists in the partially-observed observations and which encodes the variable's relationships.

Correlation matrices arising from altered input data must be not only real symmetric, but also positive semidefinite. This is an absolute requirement: even if the new, estimated correlation matrices are believed to be econometrically reliable, they may not be mathematically feasible (Rebonato & Jäckel, 1999). Once missing data were inserted in the time series data by whichever technique, correlation matrices were calculated and tested for real-symmetry and positive semi-definiteness (non-negative eigenvalues) using the technique described in Rebonato and Jäckel (1999). All correlation matrices were found to satisfy the positive semi-definite requirement.

3. Results

3.1 FX Rates

FX rates are highly liquid and therefore do not have many stale data, nor many missing data.

Correlation matrices were compared with each other according to the approach detailed in Section 3. The techniques were first tested using time series with 0.5% and 1% missing data. These were created manually: empirical data were used, and random data points removed until the requisite level of missingness was reached.

Differences between the difference matrices using the Quandl dataset for each approach are shown in Figure 4. To standardise the results, the best technique (for both 0.5% and 1.0% missing data, this was the Stine interpolation) was rebased to 100. This allows a percentage comparison to be made with results obtained from other methods. The techniques that performed the worst were the mode, random, mean, and median interpolation > 1 000% higher than the results obtained for the Stine interpolation.

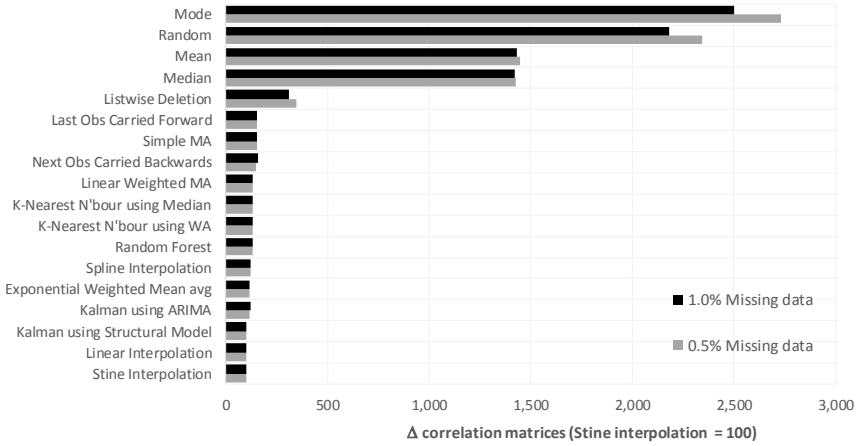


Figure 4. Comparing techniques using correlation matrix differences (for 0.5% and 1.0% missing data) for Quandl FX rates

Each technique was evaluated using other tests of imputation accuracy, namely the RMSE, the MAE and the MAPE when the quantity of missing data was 0.5% and 1.0% (still with the Quandl data set). The results, which are again presented such that the best performing technique is rebased to 100 (again, the Stine interpolation), appear in Figure 5. The moving average and random forest approaches fare the worst for all imputation methods at both levels of missing data.

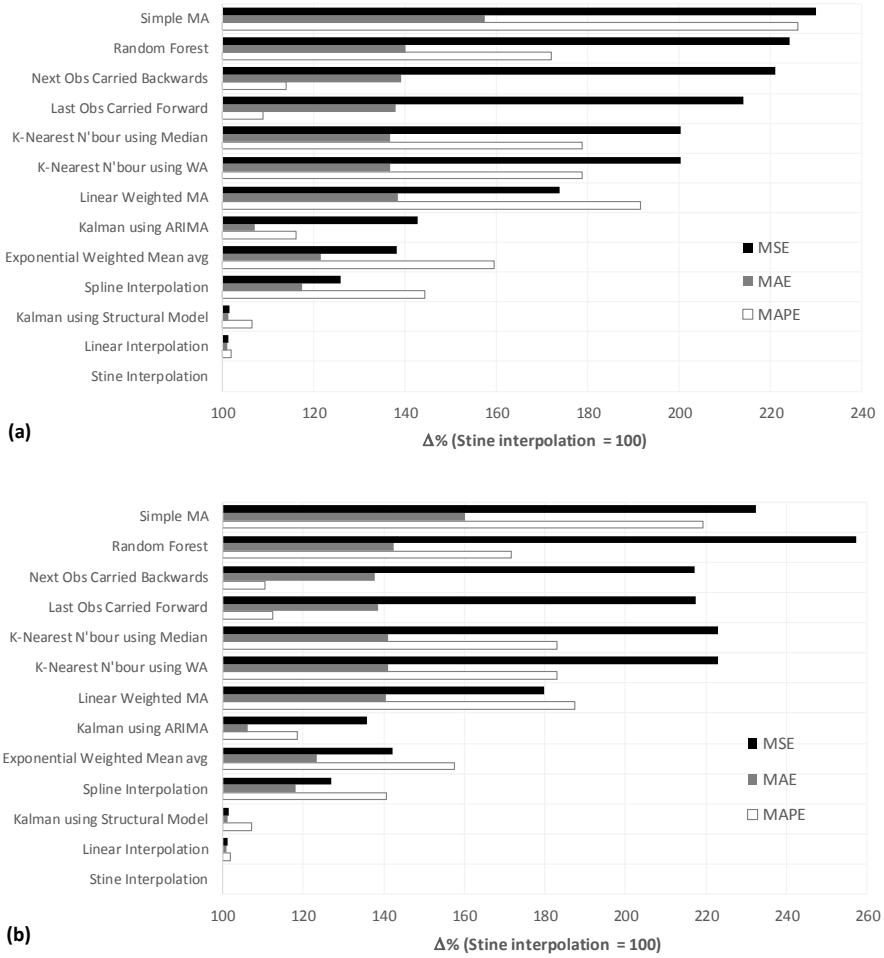


Figure 5. Comparing techniques (a) 0.5% missing data and (b) 1.0% missing data

Using the Bloomberg dataset, differences between the difference matrices for each approach are shown in Figure 6. Results are again standardised with the best technique (for both 0.5% and 1.0% missing data, this was the Stine interpolation) rebased to 100. The techniques that performed the worst were again the mode, random, mean, and median interpolation which gave difference between difference matrices > 1 500% higher than the results obtained for the Stine interpolation.

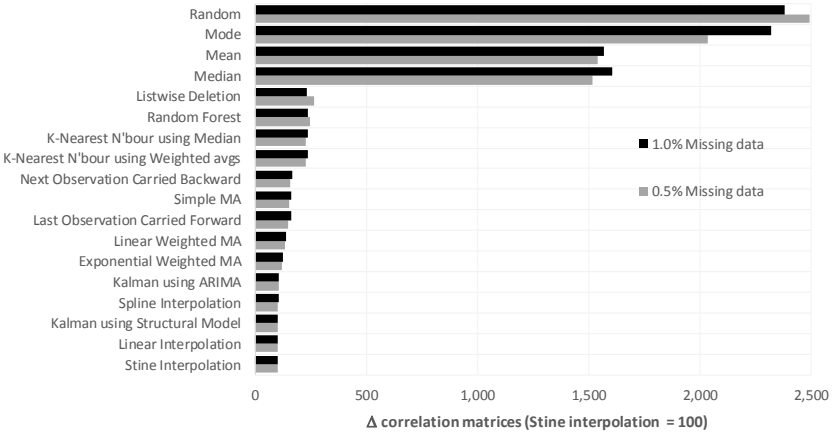


Figure 6. Comparing techniques using correlation matrix differences (for 0.5% and 1.0% missing data) for Bloomberg FX rates

Each technique was then evaluated using the imputation accuracy tests when the quantity of missing data was 0.5% and 1.0% (using the Bloomberg data set). The results appear in Figure 7, again with the best performing technique rebased to 100 (again, the Stine interpolation). The random forest approach fares the worst for all approaches at both levels of missing data.

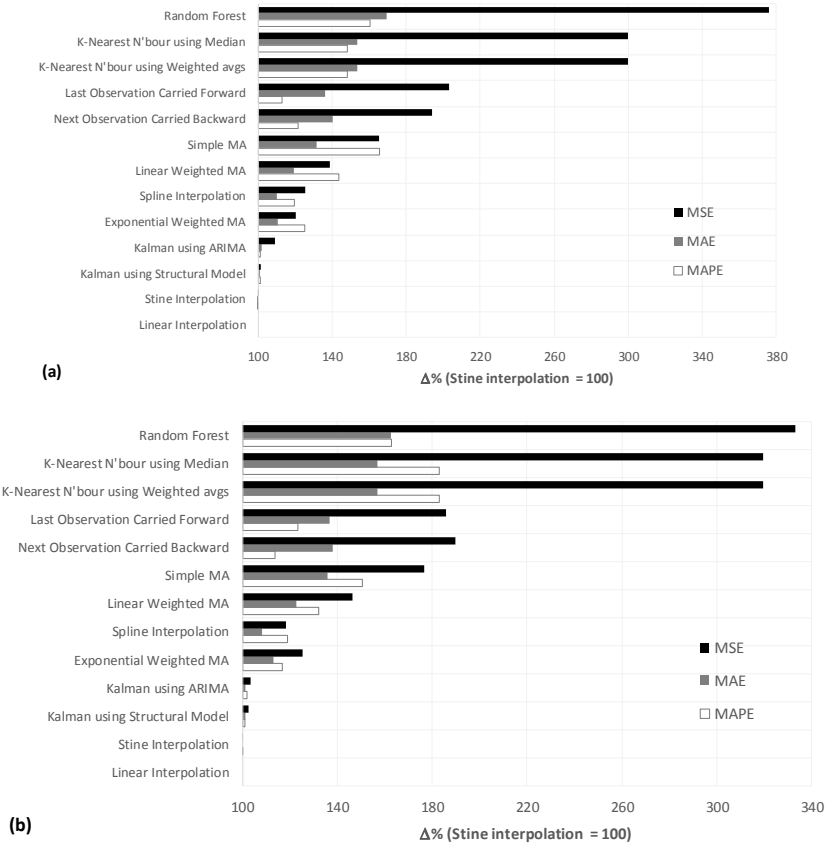


Figure 7. Comparing techniques (a) 0.5% missing data and (b) 1.0% missing data

3.2 Interest Rates

The Bloomberg dataset was used for testing the different techniques to handle missing data present in interest rates. Interest rates are generally not highly-liquid and are characterised by many stale and missing data (in the range 2% to 45%). Because of this, higher thresholds of missing (or stale) data were used, namely, 2%, 15%, 30% and 45% missing data. These levels correspond to observed missing data percentages.

The three most liquid interest rates (3m EURIBOR, 3m GBP LIBOR and 3m USD LIBOR) were used as the benchmark to test the imputation techniques: these also had the least missing data.

Differences between the difference matrices for each approach for interest rate data are shown in Figure 8. To standardise the results, the best technique (for all levels of missing data, this was the Stine interpolation) was rebased to 100. The techniques that performed the worst were (again) the mode, random, mean, and median interpolation > 1 000% higher than the results obtained for the Stine interpolation.

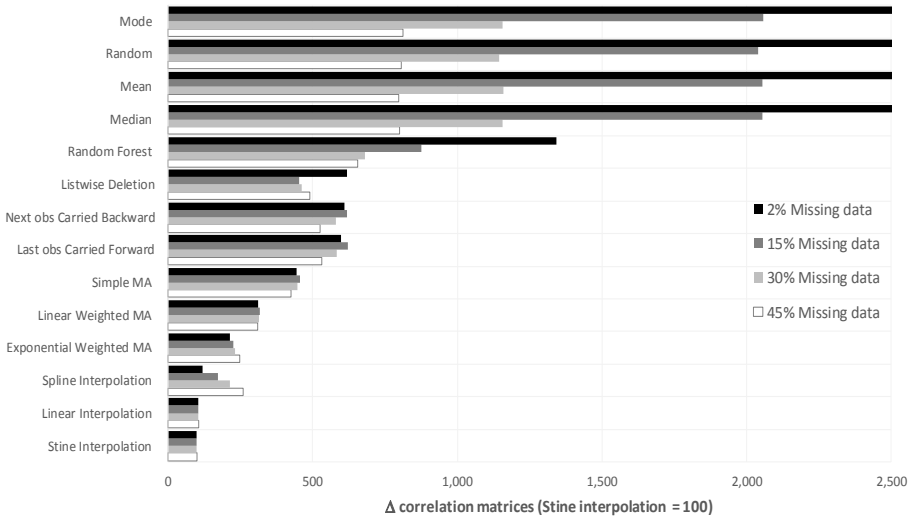
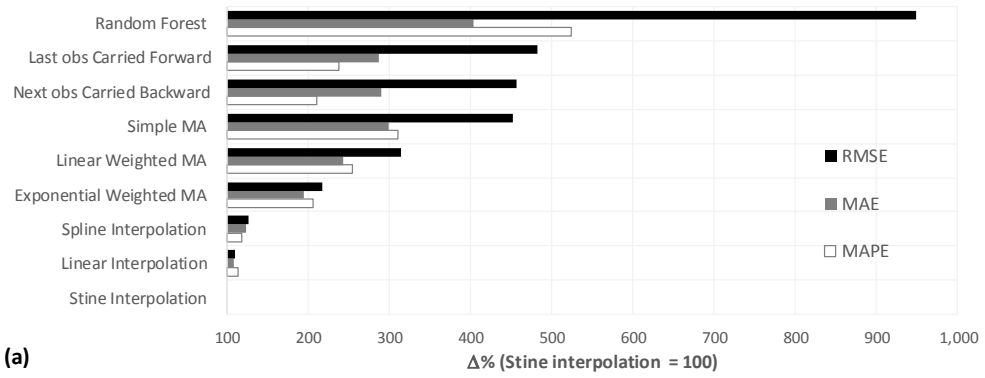


Figure 8. Comparing techniques using correlation matrix differences (for 2%, 15%, 30% and 45% missing data) for Bloomberg IR rates

The better techniques were then selected and evaluated using other tests of imputation accuracy, namely the RMSE, the MAE and the MAPE when the quantity of missing data was 2%, 15%, 30% and 45%. Results, which are again presented such that the best performing technique is rebased to 100 (again, the Stine interpolation), appear in Figure 9. The random forest method again fares the worst for all imputation methods at all levels of missing data.



(a)

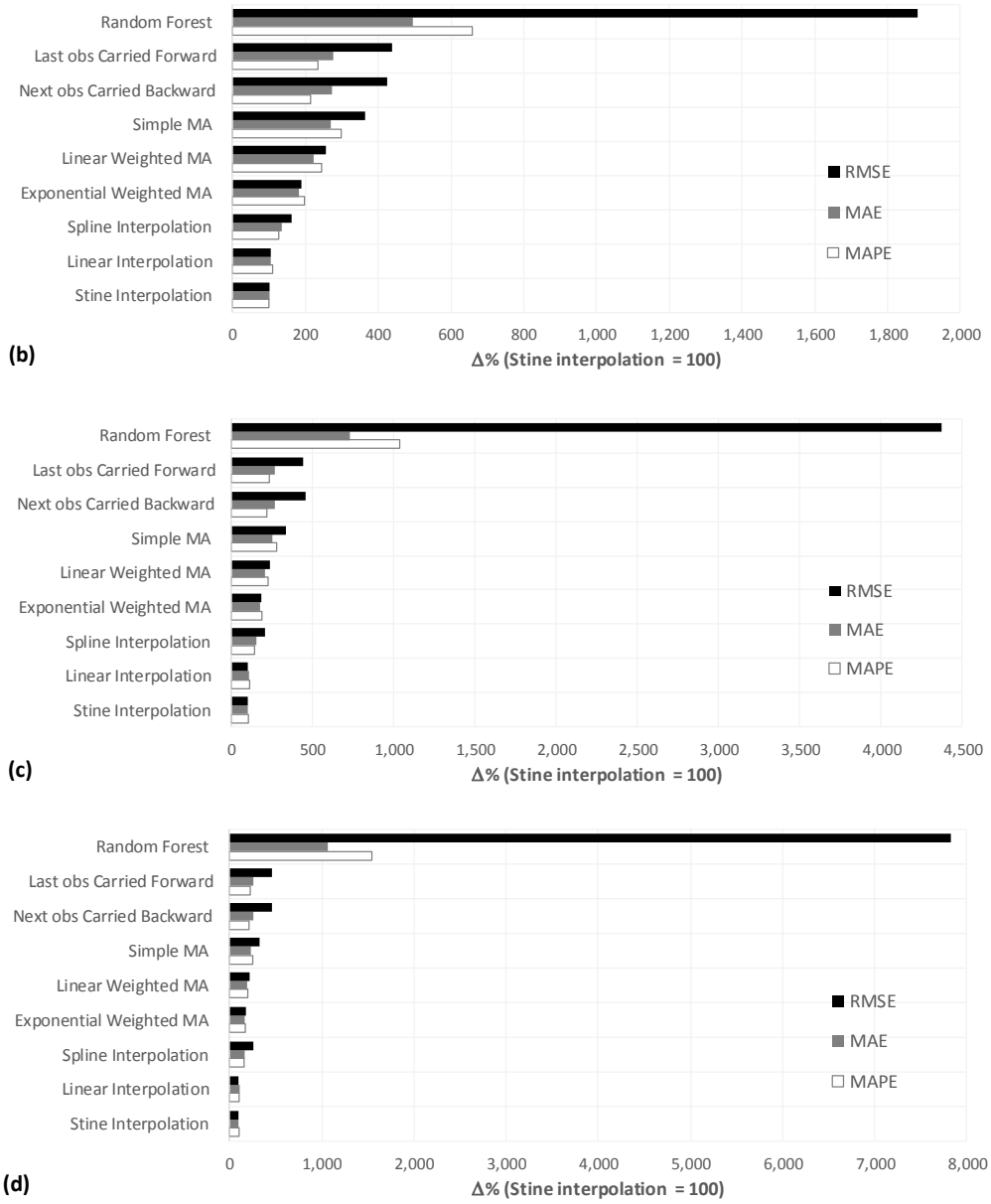


Figure 9. Comparing techniques (a) 2% (b) 15%, (c) 30% and (d) 45% missing data

4. Conclusion

The results provide a clear indication of the best and worst methods for missing data imputation and may be used for further research on different techniques and methods. There are, however, novel issues which have been set aside for future research.

Techniques to handle missing data were suggested and implemented. These techniques were used to construct correlation matrices to see which techniques created the correlation matrix that were the most accurate, when compared to the correlation matrix of a complete dataset. For FX and interest rate data, the mean, median, mode, random and listwise deletion methods were the worst performing methods.

From the RMSE, MAE and the MAPE results the Stine and linear interpolation methods are the best techniques to deal with missing data in interest rate indices. Spline interpolation and the Exponentially Weighted Moving Average methods may also be used as both these give good results. The worst performing methods are – across all levels of missingness – the random, mean, median and mode imputation approaches. These should be avoided.

The flexibility of multiple imputation methods provides a considerable advantage over methods that are reliant on the underlying data distribution. Although these methods are difficult to implement, using outputs and comparing these to univariate imputation methods can be of substantial interest.

Imputation methods such as random forest and the k-nearest neighbour use parameters inputted by the user. These parameters have not been tested for optimality – nor calibrated – in this work. Future work could establish such calibration and optimality. The parameters obtained could also be calibrated to determine which techniques are the best for use with FX, interest rates, and other time series data.

The techniques that performed the best to handle missing data in FX rate data were the Stine interpolation, linear interpolation and the Kalman method using a structural model. The techniques that performed the best to handle missing data in interest rate index data were Stine interpolation, linear interpolation and, to a lesser extent, the Exponentially Weighted Moving Average and the spline interpolation imputation methods.

Acknowledgments

This work is based on research supported in part by the Department of Science and Technology (DST) of South Africa. The grant holder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by DST-supported research are those of the author(s) and that the DST accepts no liability whatsoever in this regard. The first author carried out this research at Deloitte and Touche as part of his six-month Business Mathematics and Informatics (BMI) industry-directed research project in partial fulfilment of the degree of Master of Science in BMI.

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks: Sage. https://doi.org/10.4135/9781412985079

Allison, P. D. (2012). *Handling missing data by maximum likelihood* (pp. 1-21). In SAS® Global Forum 2012 Conference. Orlando: SAS® Institute Inc.

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined persion rs and Fisher's z transformations. *The Journal of General Psychology, 125*(3), 245-261. https://doi.org/10.1080/00221309809595548

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation.* Hoboken: John Wiley & Sons. https://doi.org/10.1002/9780470904848

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Dicesare, G. (2006). *Imputation, estimation and missing data in finance*. Waterloo: University of Waterloo. Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods.* Oxford: Oxford University Press.

Elissavet, R. K. (2017). *Missing data in time series and imputation methods*. Samos: University of the Aegean. MSc dissertation.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Fichman, M., & Cummings, J. M. (2003). Multiple imputation for missing data: making the most of what you know. *Organizational Research Methods, 6*(3), 282-308. https://doi.org/10.1177/1094428103255532

Fung, D. S. (2006). *Methods for the estimation of missing values in time series*. Perth: Edith Cowan University. MSc dissertation.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60*(2), 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: analysis and design*. New York: Springer. https://doi.org/10.1007/978-1-4614-4018-5

Harvey, A. C., & Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association, 79*(385), 125-131. https://doi.org/10.1080/01621459.1984.10477074

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer. https://doi.org/10.1007/978-0-387-84858-7

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561-581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software, 45*(7), 1-47. https://doi.org/10.18637/jss.v045.i07

Hull, J. C. (2012). *Option, Futures and other Derivatives*. Harlow: Pearson Education.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software, 27*(3), 1-22. https://doi.org/10.18637/jss.v027.i03

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, 70*(1), 1-31. https://doi.org/10.18637/jss.v070.i01

Keener, R. W. (2010). *Theoretical Statistics: Topics for a Core Course*. New York: Springer. https://doi.org/10.1007/978-0-387-93839-4

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49-69.

Kokic, P. (2001). Standard methods for imputing missing values in financial panel/time series data.

Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software, 74*(7), 1-16. https://doi.org/10.18637/jss.v074.i07

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with missing data*. Hoboken: John Wiley & Sons. https://doi.org/10.1002/9781119013563

Moritz, S., & Bartz-Beielstein, T. (2017). Time series missing value imputation in R. *The R Journal, 9*(1), 207-218.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. New York.

Oh, S. (2015). *Multiple imputation on missing values in time series data*. Durham: Duke University. MSc dissertation.

Raghunathan, T. E., Lepkowski, J. M., Soleberger, P., & van Hoewyk, J. (2001). A multivariate technique for multiply imputing missing variables using a sequence of regression models. *Survey Methodology, 27*(1), 85-89.

Rebonato, R., & Jäckel, P. (2011). *The most general methodology to create a valid correlation matrix for risk management and option pricing purposes.* Retrieved from https://ssrn.com/abstract=1969689

Rice, J. A. (2007). *Mathematical statistics and data analysis.* Duxbury.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1978). Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 30-34).

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489. https://doi.org/10.1080/01621459.1996.10476908

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research, 8*(2), 1625-1657.

Schafer, J. L. (1997). *Analyisis of incomplete multivariate data.* Boca Raton: Chapman & Hall/CRC. https://doi.org/10.1201/9781439821862

Schafer, J. L. (1998). Multiple imputation for multivariate missing-data problems: A data analyst`s perspective. *Multivariate Behavioral Research, 33*(4), 545-571. https://doi.org/10.1207/s15327906mbr3304_5

Taylor, S. J. (2008). *Modelling financial time series*. New York: World Scientific Publishing.

Tichy, T. (2006). Foreign exchange rate modelling. In: International Conference on Risk Management and Modelling. *Ostrava,* 372-380.

Trueman, C. N. (2016). The History Learning Site. Retrieved from http://www.historylearningsite.co.uk/sociology/research-methods-in-sociology/longitudinal-studies/

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064. https://doi.org/10.1080/10629360600810434

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Mulitvariate Imputation by Chained Equation in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

Yoon, J., Zame, W. R., & van der Schaar, M. (2017). *Multi-directional recurrent neural networks: A novel method for estimating missing data*. In: ICML 2017 Time Series Workshop (pp. 1-5). Sydney.

Yucel, R. M. (2011). State of the multiple imputation software. *Journal of Statistical Software, 45*(1), 1-7. https://doi.org/10.18637/jss.v045.i01

Notes

Note 1. Listwise deletion (also Complete Case Analysis) is a deletion method where any record/observation or case is excluded from an analysis if any single observation in the record is missing (Graham, 2012).

Note 2. "The cases with data for a variable, and cases with missing data for a variable, are each random samples of the total" (Graham, 2012).

Note 3. Missing does not depend on the values of the data Y, missing or observed (Little & Rubin, 2002).

Note 4. R is free programming software without warranty.

Note 5. The Pearson correlation statistic measures the strength of a linear relationship (Rice, 2007).

Note 6. An exchange rate defines the ratio of which one currency can be exchanged for another currency at any given point of time.

Note 7. An interest rate index serves as a benchmark used to calculate the interest rate charged on financial products.