**Revue Française d'Economie et de Gestion**
ISSN : 2728- 0128
Volume 6 : Numéro 4

# Handling missing data in Burundian sovereign bond market

# Imputation des données manquantes sur le marché des obligations souveraines du Burundi

**IRAKOZE Irene**
Lecturer
Actuarial Department, Institute of Applied Statistics, University of Burundi, Bubanza, Burundi
Laboratoire universitaire de recherche en modélisation et en Ingénierie de Statistique Appliquée Doctoral School, University of Burundi, Bujumbura, Burundi

**NTAWIRATSA Rédempteur**
Associate Professor
Department of Economics, Faculty of Economics and Management, University of Burundi, Bujumbura
Centre Universitaire de Recherche pour le Développement Economique et Social

**NIYUKURI David**
Lecturer
Department of Mathematics, Faculty of Science, University of Burundi, Bujumbura, Burundi
Centre de Recherche en Mathématiques et Physique

**Date de soumission** : 24/02/2025

**Date d'acceptation** : 04/04/2025

**Pour citer cet article**:

IRAKOZE. I. & AL. (2025) "Handling missing data in Burundian sovereign bond market" Revue Française d'Économie et de Gestion, « Volume 6 : Numéro 4 » pp: 461- 472.

**Abstract**

The aim of this article is to determine the best approach for filling in the missing data when constructing the yield curve for the Burundi bond market. In this paper, we explore the limitations and data availability constraints specific to the Burundian sovereign market and propose robust methodologies to effectively handle missing data. The results indicate that the Linear Regression method, and the Previous value method perform consistently well across variables, approximating a normal distribution for the error values. The non-parametric Missing Value Imputation using Random Forest (miss- Forest) method performs well for coupon rates but poorly for bond prices, and the Next value method shows mixed results. Ultimately, the Linear Regression (LR) method is recommended for imputing missing data due to its ability to approximate normality and predictive capabilities. However, filling missing values with previous values has high accuracy; thus, it will be the best choice when we have less information to be able to increase accuracy for LR.

**Keywords**: Burundian Yield curve; missing data; Linear Regression method; Previous value method; Random Forest method.

**Résumé**

L'objectif de cette recherche est de trouver la meilleure approche pour remplir les données manquantes lors de la construction de la courbe des taux d'intérêt du marché obligataire burundais. Dans cet article, nous explorons les limites et les contraintes de disponibilité des données spécifiques au marché obligataire burundais et proposons des approches robustes pour traiter efficacement le problème des données manquantes. Les résultats indiquent que la méthode de régression linéaire et la méthode de la valeur précédente donnent de bons résultats pour toutes les variables, avec une distribution normale pour les valeurs d'erreur. La méthode non paramétrique d'imputation des valeurs manquantes à l'aide de Random Forest (miss- Forest) donne de bons résultats pour les taux de coupon, mais de mauvais résultats pour les prix des obligations, et la méthode de la valeur suivante donne des résultats mitigés. En définitive, la méthode de régression linéaire (LR) est recommandée pour l'imputation des données manquantes en raison de sa capacité à se rapprocher de la normalité et de ses capacités prédictives. Cependant, le remplissage des valeurs manquantes par des valeurs antérieures est très précis; c'est donc le meilleur choix lorsque nous disposons de moins d'informations pour augmenter la précision de la méthode de régression linéaire.

**Mots-Clés:** bon du trésor; obligation du trésor; marché obligataire souverain; courbe des taux d'intérêt; données manquantes.

## Introduction

For more than two decades, Burundi, through its central bank, the Bank of the Republic of Burundi (BRB), has been issuing treasury securities Banque de la République du Burundi, 2017, Banque de la République du Burundi, 2018. It started with short-term securities called treasury bills, with a maturity of 13, 26 and 52 weeks. Issuance has been generally regular, every Wednesday of the week. Demand was mainly from commercial banks. For a little less than a decade, the maturity has been progressively extended, even to the current 10 years. Note that when the maturity exceeds one year, the securities are called bonds. The timing of bond issuance is identical to that of treasury bills, but the range of maturities is far from complete. In countries with tight and illiquid financial market, the issuance is not regular. Besides that, when issuing treasury securities, some maturities are absent from the results of the auctions either because they are evaded from the offer, or because the prices proposed by the issuer do not meet with the approval of the demand. The result is an unbalanced database in terms of maturity. This imbalance is a major handicap when it comes to calculating the yields of zero-coupon bonds, which are a prerequisite for the construction of the yield curve (Voloshyn I., 2015). The construction of an accurate yield curve is of paramount importance in evaluating financial instruments and analyzing market trends, particularly in the context of the bond market. However, the presence of missing data in the Burundian sovereign bond market poses a significant challenge to constructing a reliable yield curve. This paper aims to address this issue by exploring methodologies to effectively handle missing data and construct a robust yield curve for the Burundian sovereign bond market.

The Burundian sovereign bond market plays a vital role in the country's economy, providing a platform for capital mobilization, investment, and risk management. However, the unique dynamics and challenges of this emerging market present distinct obstacles to yield curve construction. Inadequate data availability and limitations in data collection practices contribute to missing data, which in turn hinders accurate yield curve estimations.

The objective of this research is to propose robust methodologies that mitigate the impact of missing data on yield curve estimations in the Burundian sovereign bond market. By employing suitable statistical techniques, we aim to enhance the transparency, reliability, and efficiency of the market by providing more accurate yield curve estimations. This, in turn, will aid investors, market participants, and policymakers in making informed decisions regarding financial instruments, investment strategies, and risk management. To achieve our objective, we employ a range of statistical techniques tailored to the

characteristics and limitations of the Burundian sovereign bond market.
These methodologies will enable us to handle missing data effectively and construct a reliable yield curve.
Ultimately, this paper addresses the issue of handling missing data in the context of constructing a reliable yield curve for the Burundian financial bond market. By proposing robust methodologies and employing suitable statistical techniques, we aim to mitigate the impact of missing data and provide more accurate yield curve estimations. The outcomes of this research have the potential to enhance transparency, reliability, and efficiency in the Burundian sovereign bond market, facilitating informed decision-making for investors, market participants, and policymakers, while contributing to the overall development of the financial sector in Burundi.
The bootstrapping technique has been used to construct the zero coupon yield curve that reflect the actual daily yield movements in the bond market (Hagan, P. S. & West, G., 2008). However, for this to be possible, we must have a complete series of zero coupon rates, prices of bonds and the corresponding maturities what is not the case for the Burundian bond market due to its illiquid nature.
The construction of an accurate yield curve is crucial for analyzing and valuing financial instruments in the bond market. The Burundian sovereign bond market, data availability is limited while the presence of missing data poses a significant challenge in accurately constructing the yield curve. This work aims to address the issues related to missing data and its impact on constructing a reliable yield curve for the Burundian sovereign bond market.
The Burundian bond market serves as a vital platform for issuing and trading fixed-income securities, providing an avenue for raising capital for both government and corporate entities. However, due to factors such as data collection limitations, reporting delays, incomplete data records, the avail- ability of historical bond market data in Burundi is insufficient and contains gaps. Missing data can occur in various forms, including absent observations for specific time periods, incomplete information on bond characteristics, or insufficient transaction data.
The presence of missing data poses challenges when attempting to construct a yield curve, which is a graphical representation of the relationship between bond yields and their respective maturities. An accurate yield curve is crucial for estimating market interest rates, valuing bonds, and conducting risk assessments (Evans C. L. & Marshall D. A., 2007). However,

if missing data is not appropriately handled, it can lead to biased or inaccurate yield curve estimations, potentially impacting investment decisions, risk management strategies, and overall market efficiency.

The problem at hand requires the development of robust methodologies and techniques to handle missing data effectively in the construction of the yield curve for the Burundian financial bond market. The proposed solution considers the unique characteristics of the market, data availability constraints, and statistical techniques suitable for addressing missing data issues. Addressing this problem contributes to enhancing the transparency, reliability, and efficiency of the Burundian bond market. Providing accurate yield curve estimations despite the presence of missing data allows market participants, investors, and policymakers to have access to more reliable in- formation for pricing bonds, assessing risk, and making informed investment decisions. Moreover, an improved understanding of the yield curve dynamics will aid in the development of financial products, trading strategies, and overall market.

The rest of the article covers, the methodology, the results, the discussion of the results and the conclusion.

## 1. Methodology

Numerous methods have been devised in the literature to calculate the yield curve. Nevertheless, when confronted with the task of computing the yield curve through the bootstrapping technique, a challenge arises in dealing with missing data. These missing data points pertain to maturities that were not issued or did not attract buyers, necessitating an imputation process. Missing data mechanisms have implications on methods to handle the missing data (Nakagawa S., 2015). Mainly, we have three types of missingness mechanisms:

- Missing completely at Random (MCAR): where missingness of a variable is completely independent of itself and other variables. In other words, the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself.
- Missing not at Random (MNAR): where missingness of a variable is related to itself.
- Missing at Random (MAR): where missingness of a variable is dependent on another variable. The probability of missing data on a variable is related to some other measured variable in the model, but not to the value of the variable with missing

values itself.

For Burundi financial market, the missingness of some maturities is due to mainly when there is no one to buy the bonds, or if there is no issuance which is often influenced by the treasury needs. Thus, to encompass those data missingness scenarios, we considered the missingness data, to be missing completely at random.

To handle missing data, various imputation methods from the literature are explored: García-Laencina, P. J., et al. (2009), Josse, J. & Husson, F. (2012), Malarvizhi, R. & Thanamani, A. S. (2012). The K nearest neighbor (KNN) approach involves selecting the K nearest observations (neighbors) based on a distance metric when dealing with an incomplete pattern. These chosen observations have known values for the features that need to be imputed. To estimate each incomplete feature value, a weighted average of these known values is calculated. Also, the model does not involve the creation of explicit predictive models as it relies on the training data-set itself as a" lazy" model. The methods available for handling missing values in principal component analysis (PCA) primarily offer point estimates of the parameters, such as axes and components, as well as estimates of the missing values. These methods provide single-value approximations rather than providing a range or distribution of potential values for the parameters or missing values. I this work, the following methods are considered for filling in the missing values:

- Fill in missing values with previous or next value (with *fill* function of **tidyr** package in R)
- Fill in missing values with the values of the nearest neighbors (with *knnImputation* function of **DMwR** package in R)
- Fill in missing values with Multivariate Imputation by Chained Equations, known as MICE (Schmitt P., Mandel, J. Guedj, M., 2015) (with *mice* function of **mice** package in R)
- Non-parametric Missing Value Imputation using Random Forest (with *missForest* function of **miss Forest** package in R)
- Linear regression with same maturity data issued previously.

To test these methods, a data set of financial transactions with complete maturities is utilized. A random subset of the data is chosen to introduce missing maturities, based on the proportion of financial data with complete maturities from Burundi Central Bank, which stands at 65% (which means, we had 35% of missing data). Additionally, missing values for price and yield are generated. The data-set containing the created missing data is then employed to apply the six imputation methods. Following the imputation process, the accuracy of each method is assessed by calculating the Mean Absolute Error (MAE) between simulated and observed data. This computation is repeated multiple times to determine the distribution of error values for each method. Finally, the Shapiro test is conducted to examine whether the error values are normally distributed. If the $p - value > 0.05$, it implies that the distribution of the data are not significantly different from normal distribution.

## 2. Results

The following presents the results, showcasing the error values obtained from various models. These results enable us to assess the performance of each model concerning data imputation for the Burundian bond market. Based on Figure 3, it appears that utilizing the previous values to fill the missing data results in low error values for the bond prices. This approach is closely followed by the implementation of linear regression.

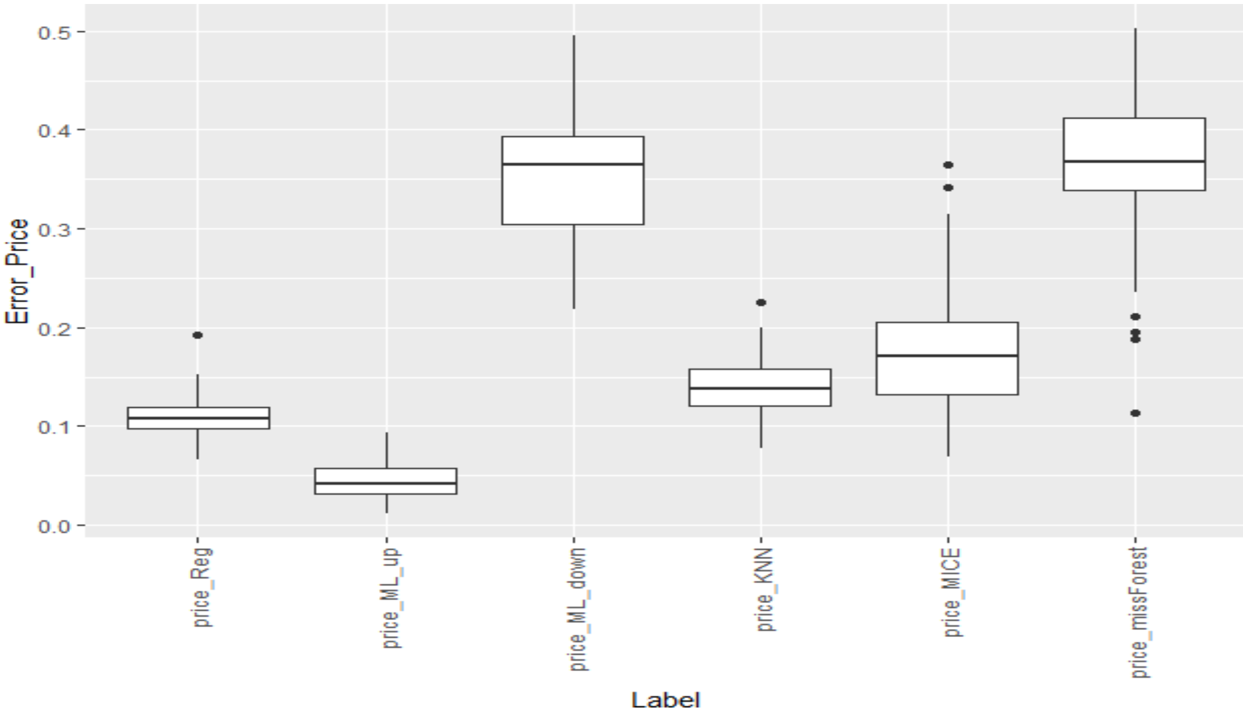


**Figure 1: Box-plots of the mean absolute error values computed for bonds prices after data imputation with six methods.**

According to Figure 2, the error values for the bond prices show a similar trend of being low when employing the previous values for filling the missing data. Linear regression follows this method closely.

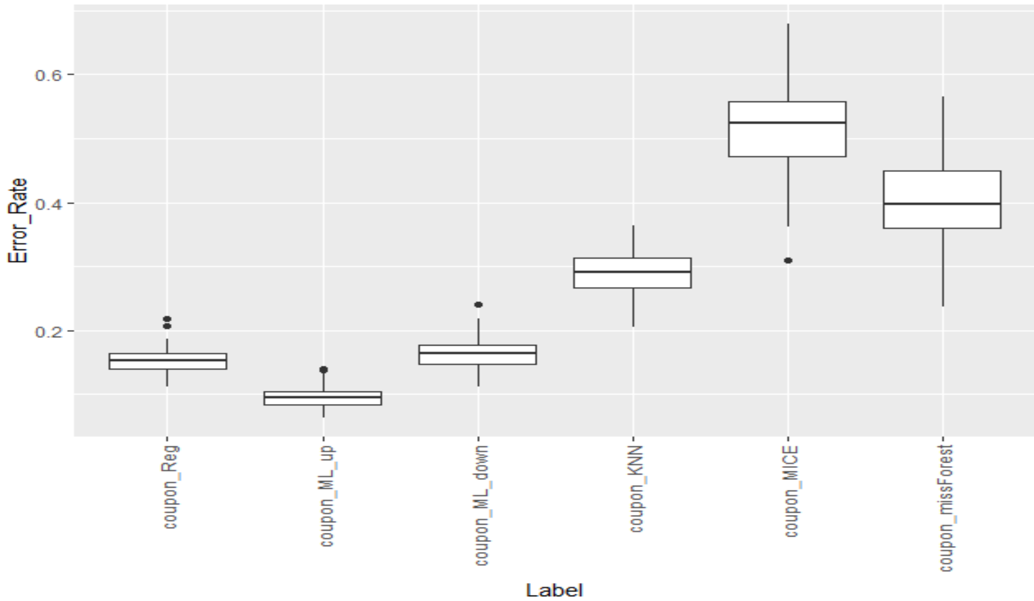


**Figure 2: Box plots of Mean absolute Error values computed for coupon rates after data imputation with six methods Based on Figure 3, it can be observed that utilizing the previous values to fill the missing data leads to low error values for the bond prices. This approach is closely followed by the implementation of linear regression.**

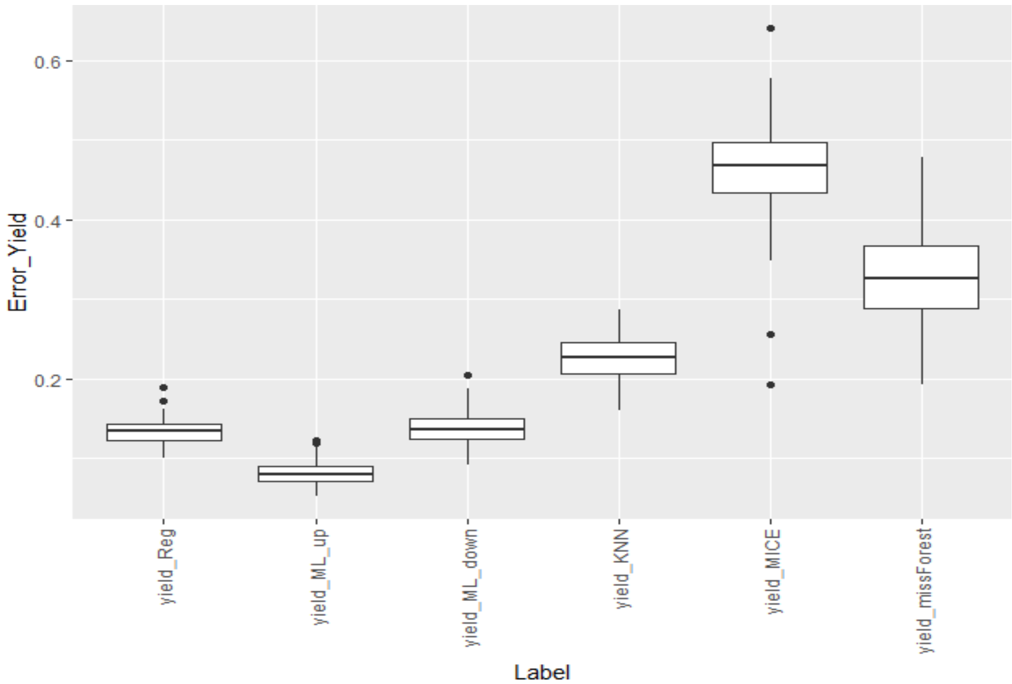


**Figure 3: Box plots of Mean absolute Error values computed for the yield after data imputation with six methods**

In all three cases, employing the method of filling the missing data with previous values demonstrated superior performance compared to other methods. Table 1 illustrates that, overall, the error values for bond prices, yields, and coupon rates exhibited a normal distribution.

**Table 1: Normality test for mean absolute error values computed after data imputation**

| Method | Estimate | p-value |
|---|---|---|
| | Coupon | 0.7794 |
| Linear Regression | Price | 0.0666 |
| | Yield | 0.8639 |
| | Coupon | 0.8991 |
| Miss Forest | Price | 0.0009 |
| | Yield | 0.5212 |
| | Coupon | 0.4628 |
| MICE | Price | 0.0171 |
| | Yield | 0.9425 |
| | Coupon | 0.6632 |
| KNN | Price | 0.4116 |
| | Yield | 0.1631 |
| | Coupon | 0.9530 |
| Previous value | Price | 0.0053 |
| | Yield | 0.5682 |
| | Coupon | 0.0880 |
| Next value | Price | 0.8073 |
| | Yield | 0.1118 |

## 3. Discussion

Based on the results of the normality tests for mean absolute error (MAE) values computed after data imputation, we can compare the different methods in terms of how well they approximate a normal distribution for the error values. A higher p-value (greater than 0.05) indicates better adherence to the normality assumption. We compare the methods for each variable in the following subsections.

### 3.1. Coupon rates

The results of coupon rates show that for

- Linear Regression (p-value: 0.7794), the error values for coupon rates are likely to follow a normal distribution.
- Miss Forest (p-value: 0.8991), the error values for coupon rates are highly likely to follow a normal distribution.
- MICE (p-value: 0.4628), the error values for coupon rates are likely to follow a normal distribution.
- KNN (p-value: 0.6632), the error values for coupon rates are likely to follow a normal distribution.

- Previous value (p-value: 0.9530), the error values for coupon rates are highly likely to follow a normal distribution.
- Next value (p-value: 0.0880): The error values for coupon rates are not following a normal distribution.

For the coupon rate, the previous value method performs the best in terms of adhering to the normality assumption for the error values in the coupon rate variable. The miss Forest method also shows good results.

### 3.2. Price

The results of price of bonds show that for

- Linear Regression (p-value: 0.0666), the error values for bond prices are approximately normally distributed.
- Miss Forest (p-value: 0.0009), the error values for bond prices do not follow a normal distribution.
- MICE (p-value: 0.0171), the error values for bond prices are approximately normally distributed.
- KNN (p-value: 0.4116), the error values for bond prices are likely to follow a normal distribution.
- Previous value (p-value: 0.0053), the error values for bond prices are approximately normally distributed.
- Next value (p-value: 0.8073), the error values for bond prices are likely to follow a normal distribution.

For the bond price, we can see that the Linear Regression method and the previous value method both perform well in terms of normality for the error values in the price variable. The miss Forest method, on the other hand, shows the poorest adherence to the normality assumption for this variable.

### 3.3. Yield

The results of yield from the bonds show that for

- Linear Regression (p-value: 0.8639), the error values for yields are highly likely to follow a normal distribution.
- Miss Forest (p-value: 0.5212), the error values for yields are likely to follow a normal distribution.
- MICE (p-value: 0.9425), the error values for yields are highly likely to follow a normal distribution.
- KNN (p-value: 0.1631), the error values for yields are likely to follow a normal

distribution.

- Previous value (p-value: 0.5682), the error values for yields are likely to follow a normal distribution.
- Next value (p-value: 0.1118), the error values for yields are not following a normal distribution.

For the yield curve, we realize that the Linear Regression method and the MICE method both show strong results in adhering to the normality assumption for the error values in the yield variable. The KNN method and the previous value method also exhibit reasonably good performance.

Overall, the Linear Regression method and the previous value method consistently show better results across the variables, with higher p-values indicating better adherence to the normality assumption for the error values. The miss Forest method performs well for coupon rates but poorly for bond prices. The Next value method shows mixed results, with better performance for bond prices and poorer performance for yields.

These results argue the findings of Van Buuren, S. & Groothuis-Oudshoorn, K. (2011), which indicate that the most popular methods (KNN, MICE) are not necessarily the most efficient ones even though they did not compare it with the previous value and the linear regression method.

In conclusion, the normality tests indicate that both the Linear Regression method and the previous value method are recommended for imputing missing data in this study. These methods consistently generate error values that closely approximate a normal distribution across the variables. However, the Linear Regression method stands out as the preferred choice, not only because of its ability to approximate normality but also due to its predictive capabilities in estimating the missing data. Researchers should consider these results when selecting an appropriate imputation method for similar financial market data analyses.

**Conclusion**

The objective of this research was to determine the best imputation methods to fill the gaps observed in the Burundian primary bond market database. To achieve this, we first revisited the most commonly used imputation methods. We then tested them using data from the Burundian primary market for treasury securities. Our tests yielded the following results: the Linear Regression method and the MICE method both show strong results in adhering to the normality assumption for the error values in the yield variable. The KNN method and the previous value method also exhibit reasonably good performance.

Overall, the Linear Regression method and the previous value method consistently show better results across the variables, with higher p-values indicating better adherence to the normality assumption for the error values. The miss Forest method performs well for coupon rates but poorly for bond prices. The Next value method shows mixed results, with better performance for bond prices and poorer performance for yields.

Given that the secondary market for these securities is developing, the data from it will enrich the database from which the yield curve will be daily estimated. Like the primary market, the secondary market will certainly have gaps that will need to be filled. It will therefore be necessary to continue the search for imputation methods adapted to the nature and extent of the missing data. This is one of the avenues for future research.

## References

**Banque de la République du Burundi** (2018), Convention entre le Ministère des Finances et la BRB portant sur les Emissions et les transactions des titres du Trésor.

**Banque de la République du Burundi** (2017), Instruction No 3/2017 portant réglementation du marché des titres du trésor du Burundi

**Evans, C. L. & Marshall, D. A.** (2007), Economic determinants of the nominal treasury yield curve, Journal of Monetary Economics, 54, 1986– 2003.

**García-Laencina, P. J., et al.** (2009). K nearest neighbors with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 72(7-9), 1483-1493.

**Hagan, P.S. & West, G** (2008), Methods for constructing a yield curve, Wilmott Magazine, May, 70–81.

**Josse, J. & Husson, F**. (2012), Handling missing values in exploratory multivariate data analysis methods, Journal de la Société Française de Statistique, 153, 79–99.

**Malarvizhi, R., & Thanamani, A. S.** (2012). K-nearest neighbor in missing data imputation. Int. J. Eng. Res. Dev, *5*(1), 5-7.

**Nakagawa, S.** (2015), Missing data: mechanisms, methods and messages, Eco- logical statistics: Contemporary theory and application, 81–105.

**Schmitt, P., Mandel, J. & Guedj, M.** (2015). A comparison of six methods for missing data imputation. Journal of biometrics & biostatistics, 6(1), 1.

**Van Buuren, S., & Groothuis-Oudshoorn, K**. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, *45*, 1-67.

**Voloshyn, I.** (2015), Direct extracting the forward yield curve from the coupon bond prices, Available at SSRN 2499463