International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

Data Imputations for WWTPs Quality Variables: case study of Zaio's WWTP (Morocco)

Abdellah CHAOUI [1], Wafae EL KHOUMSI [1], Mohamed LAAOUAN [2], Rqia BOURZIZA [1], Karima SEBARI [1]*

[1] *IAV Hassan II, Rabat, Morocco.*
[2] *International Institute for Water and Sanitation, Rabat, Morocco.*

Abstract

Complete time series variables of influent and effluent variables are essential for enhancing the uncertainty of processes within wastewater treatment plants (WWTPs). In addition to that, these time series are essential for conducting different analyses and simulations to enhance WWTPs design, operations, and control. However, it is very common within WWTPs to find missing values within these variables, which leads to replacing them using traditional approaches. For this, the current paper aims at applying 14 univariate data imputation models, including the most commonly used ones among professionals in the field, to predict missing values of a time series variable (BOD5). This contribution also provides comparison tools that aim at selecting the model that yield the least error terms based on MAE, MSE, RMSE, and R-squared. Findings indicate that the autoregressive integrated moving average (ARIMA) imputation model is the most suitable to replace missing data compared to other models, as it ranks the first in terms of the least error terms and the highest coefficient of determination.

Keywords: Data imputation, Data prediction, WWTPs, BOD5, Error terms, R-squared

Full length article **Corresponding Author*, e-mail: abdellahchaoui2000@gmail.com

1. Introduction

Wastewater Treatment Plants (WWTPs) are infrastructures existing in urban areas and generating biodegradation processes that naturally occur near rivers to remove as much suspended solids as well as contaminants from water before discharging it to nature or reusing it [2]. Many contributions indicate that the high costs of these facilities, that accounts for both operational and capital costs [1; 4], has led to the use of different simulation models that enable enhancing their performance [11] in terms of WWTP design [9], operations [5], and control [7]. Yet, one of the limitations of implementing these simulations relates to some missing experimental observations within the data sets [10], which is mainly caused by the high cost of experimental collection of these parameters. Furthermore, other reasons can be included, but are not limited to, the high cost and inefficiency of advanced technologies such as on-line sensors, special events such as holidays, and the lack of experience of personnel during some shifts [8]. It is important to note that WWTPs datasets are time series since they are recorded on consecutive periods of time to measure the variability of the plants' data, which is considered as the most important factor to limit the uncertainty within the wastewater treatment processes and systems [8]. The following contribution provides 14 different univariate data imputation models including traditional ones such as completing data by random sample as a reference, and comparing them using error terms to suggest the most accurate one. The major empirical issue addressed in this research is whether if traditional data imputation methods are reliable or should be replaced by other data replacement methods.

2. Materials and Methods

The following section will introduce the suggested univariate models used for missing data imputation. R studio was used in this research to compile models and generate predictions for the Biochemical Oxygen Demand for five days (BOD5) variable, which is a quality variable measured in all WWTPs. The original data represent the measured value of BOD5 for an Zaio, Morocco wastewater treatment plant for 90 consecutive days. In concordance with the purpose of this research, observations during week-ends were removed intentionally in order to predict them using different univariate models. This is to compare the predictions of each model to actual values using different tools that will be introduced in a later section of this paper.

2.1. Missing value imputation by random sample

The first data replacement method uses a random sample that consists of replacing missing values by random values that ranges between the minimum and the maximum values of the sample's observations. This method is very common among professionals, and while it is considered as statistically correct, it results in different types of bias when it is used to make further statistical inferences.

2.2. Missing value imputation by inferential statistics

For the second imputation method, it consists of replacing missing values by different inferential statistics that are mainly the mean, the median, or the mode of the sample observations [13]. This method is used as a reference method since it is one of the most common techniques used among professionals.

2.3. Missing value imputation by observations carried backward or forward

Within this section, two main data replacement methods are introduced. The first method is imputation using the last observation carried forward that is also referred to as LOCF. The main assumption of this method indicates that the last observed value will not change over time [15]. But for the second imputation method, which is the next observation carried backward (NOBD), it refers to the reverse direction of the LOCF method, as it duplicates the next value backward.

2.4. Missing value imputation by interpolation

Interpolation is a method of estimating missing values by relying on actual observations. The first method used in this section is the linear interpolation [14]. This method uses a linear function to find the approximation of the value of a function f(x) such as:

$$L(x) = a(x - x_k) + b \quad (1)$$

In this case, both parameters (a and b) are selected in a way that makes the values of the function L(x) in accordance with the values of the function f(x) such as:

$$L(x_1) = f(x_1) \quad (2)$$

These conditions are satisfied by the given function:

$$L(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) + f(x_1) \quad (3)$$

Which approximates within the interval $[x_1, x_2]$ the function f(x), taking into account the error by the following formula:

$$R^f = f(x) - L(x) = \frac{f^{"}(\xi)}{2}(x - x_1)(x - x_2), \xi \in [x_1, x_2] \quad (4)$$

Concerning the Spline interpolation, it consists of using piecewise polynomials to approximate the function f(x). It is different from the linear interpolation as it should satisfy further conditions such as the cubic Spline [16]. Thus, the function to approximate the Spline interpolation should satisfy the following relation:

$$\int_a^b \left[f^{(k)}(t) - S_{2k-1}^{(k)}(\Delta_n, t)\right]^2 dt = \int_a^b \left[f^{(x)}(t)\right]^2 dt - \int_a^b \left[S_{2k-1}^{(k)}(\Delta_n, t)\right]^2 dt \quad (5)$$

Where $S_3(\Delta_n, x)$ is an example of the cubic Spline where Δ_n is the partition $a = x_0 \leq x_1 \leq \cdots \leq x_n$, the two end points that are reconstructed by piecewise cubic polynomials and has a continuous second order derivatives [16].

For the Stine interpolation, it was developed by Stineman, and it is faster and more accurate from 2 and 6 times compared to the cubic spline interpolation. This method provides more or less similar results as the spline and linear interpolation, yet, the only advantage is that it is more reliable in the case where the variable presents an abrupt change in slope [3].

2.5. Missing value imputation by structural model & Kalman Smoothing and ARIMA

In this section two models are discussed that are ARIMA modeling and the Structural modeling. Concerning ARIMA, which stands for auto-regressive integrated moving average, it was first introduced by Box and Jenkin in 1976, and its general equation of successive differences at the dth difference is given such as:

$$\Delta^d X_t = (1 - B)^d X_t \quad (6)$$

Where d is the difference order. In this work, the general ARIMA (p,d,q) is expressed as:

$$\Phi_p(B) W_t = \theta_q(B) e_t \quad (7)$$

Where $\Phi_p(B)$ is an autoregressive operator of order p, $\theta_q(B)$ is a moving average operator of order q, and $W_t = \Delta \mathrm{d} X_t$.

Concerning the Kalman filter that uses structural modeling, it uses a recursive procedure for computing the optimal estimator of the state vector at time t, based on the information available at time t.

2.6. Missing value imputation by moving average

Regarding this section, the data replacement method is moving average. This method consists of replacing a specific missing value using weighted moving average values on both its sides. For instance, and in order to impute a missing value at a specific location i, the weighted average of $y_{i-2}, y_{i-1}, y_{i+1}, and\ y_{i+2}$ are used to calculate the moving average window size of 4 [6]. It is important to note that this contribution includes three types of the moving average imputation models that are [6]:

- Simple moving average: That assumes that all observations used are equally weighted.
- Linear weighted moving average: That assumes that the weights of the used observations are decreasing in an arithmetic progression.
- Exponential weighted moving average: That assumes that the weights of the used observations are decreasing exponentially.

2.7. Comparison method

This section will present the tools that will be used to compare between the actual values, and the ones resulting from each univariate model used in this contribution. The methods used are referred to as absolute prediction error that is based on the absolute error calculation based on the calculation of the value:

$$e_t = y_t - f_t^{(m)} \quad (8)$$

Where y_t is the actual value, and $f_t^{(m)}$ is the predicted value at time = t using the model m. This accounts for the mean absolute error (MAE), mean square error (MSE), and the root mean square error (RMSE); These variables are defined such as MAE is the measure of the absolute difference between the observed and the predicted values, MSE is the average squared error for the prediction values and RMSE accounts for the square root that is introduced to make scale of the errors to be the same as the scale of the target. It is important to note that the selection of the best univariate model to predict the missing values of the variable BOD5 will be based on the lowest error terms of all models. The models in this contribution will also be compared using the R-square (R2) that is also referred to as the coefficient of determination. This coefficient has a magnitude restricted between 0 and 1, which measures the proportion or the percentage of dependent variables that are explained by an independent variable. The higher the value of R-square, the better the model of data prediction fits actual values.

3. Results and Discussion

The following contribution used R in order to compile and graph the different prediction models as displayed in Appendix A. The MAE, MSE, RMSE, and R2 were used as the comparison tool for the applied models in order to compare the error terms. These latter are considered as a decision tool in the case of this contribution to assess the data imputation model that predicts the best missing values for the case of the time series variable BOD5. Comparison results were sorted based on the MAE in Table 1.

The most basic data imputation models used among professionals in the wastewater treatment industry are by random sample, least observation carried forward, next variable carried backward, and data imputation models using inferential statistics. While all of these models are statistically correct, findings indicate that they don't take part of the top 5 models in terms of least error terms, without any contradiction between all comparison tools. For the MAE, it resulted in values of at least 8.00 for these latter models, compared to a value between 5.01 and 5.99 of the top five models. Concerning the MSE, it resulted in values of at least 405.00 for traditional models compared to a value between 219.80 and 280.20 for the first five best models. This aligns with the findings of the RMSE, as this latter resulted in a value of at least 20.12 for traditional models compared to a value between 14.83 and 16.74 for the top five models.

Results indicate that traditional data imputation models are highly likely to yield a bias in further analyses, which is mainly due to the high error term, and the lower values of the coefficients of determination (R2) compared to other univariate models. E.g. the R2 of the random sample indicates that only 41% of the variations are explained by the independent variable, that is the actual time series variable (BOD5).

Table1: MAE, MSE, RMSE, and R2 using different imputation **models.**

Time series model	MAE	MSE	RMSE	R2
ARIMA	5,01	219,80	14,83	0,96
Structural model	5,04	221,23	14,87	0,96
Simple moving average	5,36	236,68	15,38	0,95
Linear moving average	5,65	244,91	15,65	0,95
Exponential moving average	5,99	280,20	16,74	0,94
Linear interpolation	7,17	512,78	22,64	0,90
Data imputation using the mode	8,00	405,00	20,12	0,91
Least observation carried forward	8,25	471,25	21,71	0,91
Stine interpolation	8,25	582,70	24,14	0,89
Data imputation using the median	9,50	490,00	22,14	0,89
Next observation carried backward	12,50	1371,25	37,03	0,75
Data imputation using the mean	13,15	826,30	28,75	0,82
Spline interpolation	16,24	1566,17	39,57	0,72
Random Sample	30,65	5716,90	75,61	0,41

Comparison tools indicate that among the data imputations models applied in this contribution, ARIMA is the most suitable one for predicting and replacing data, followed by the structural model, simple moving average, linear moving average, and then, exponential moving average.

4. Conclusions

This paper puts emphasis on the importance of having complete datasets of influent and effluent data in WWTPs that will enable conducting further analyses, simulations, and forecasting to enhance the efficiency, the monitoring, and the control of these plants. Measuring experimental values with high frequencies (e.g. daily basis) always enhances the quality of WWTPs' monitoring and control. Yet, in real life situations, professionals measure quality variables (e.g. BOD5) only twice or three times per week, which limits analyzing the plants' uncertainty, or conducting further analyses. In some cases, this latter situation urges professionals to complete different time series variables using basic and simple univariate data imputation models such as replacing missing values with random variables between the minimum and the maximum values, or duplicating the last observed values. Thereby, the current contribution compares between 14 different univariate data imputation models and includes the commonly used ones among professionals as a reference. In the context of this study, results show that ARIMA is the most suitable univariate model for data replacement. This is because it was ranked first with regards to models that have the least error terms using MAE, MSE, and RMSE. Furthermore, this model also has a high coefficient of determination that indicates that more than 96% of the total variations are explained by the independent variable. Additionally, it is important to note that data imputation using a structural model or a simple moving average also have low error terms, and can also be applied to predict missing data.

References

[1] C. Liu, L. Shuai, and F. Zhang, The oxygen transfer efficiency and economic cost analysis of aeration system in municipal wastewater treatment plant, Energy Procedia 5 (2011), 2437-2443.

[2] C. Martin, and A. P. Vanrolleghem, Analysing, completing, and generating influent data for WWTP modelling: a critical review, Environmental Modelling & Software 60 (2014), 188-201.

[3] G. M. Perillo, and M. C. Piccolo, An interpolation method for estuarine and oceanographic data, Computers & Geosciences 17.6 (1991), 813-820.

[4] G. Rodriguez-Garcia, M. Molinos-Senante, A. Hospido, F. Hernández-Sancho, M. T. Moreira, & G. Feijoo, Environmental and economic profile of six typologies of wastewater treatment plants, Water research 45.18 (2011), 5997-6010.

[5] G.S. Ostace, V. M. Vasile, and P. Ş. Agachi, Cost reduction of the wastewater treatment plant operation by MPC based on modified ASM1 with two-step nitrification/denitrification model, Computers & chemical engineering 35.11 (2011), 2469-2479.

[6] H. Hassani, M. Kalantari, and Z. Ghodsi, Evaluating the performance of multiple imputation methods for

handling missing values in time series data: A study focused on East Africa, Soil-Carbonate-Stable Isotope Data, Stats 2.4 (2019), 457-467.

[7] I. Nopens, L. Benedetti, U. Jeppsson, M. N. Pons, J. Alex, J. B. Copp, P. A. Vanrolleghem, Benchmark Simulation Model No 2: Finalisation of plant layout and default control strategy, Water Science and Technology 62.9 (2010), 1967-1974.

[8] J. Huo, C. Cox, W. Seaver, B. Robinson, & Y. Jiang, Innovative Missing Data Replacement Methods Using Time Series Models, World Environmental and Water Resources Congress 2008, (2008), doi:10.1061/40976(316)670.

[9] L. Benedetti, B. De Baets, I. Nopens, & P. A. Vanrolleghem, Multi-criteria analysis of wastewater treatment plant design and control scenarios under uncertainty, Environmental modelling & software 25.5 (2010), 616-621.

[10] L. Rieger, I. Takács, K. Villez, H. Siegrist, P. Lessard, P. A. Vanrolleghem, & Y. Comeau, Data Reconciliation for Wastewater Treatment Plant Simulation Studies—Planning for High-Quality Data and Typical Sources of Errors, Water environment research 82.5 (2010), 426-433.

[11] M. Henze, W. Gujer, T. Mino, M. Van Loosdrecht, Activated sludge models ASM1, ASM2, ASM2d and ASM3. IWA publishing (2000).

[12] M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, & V. A. E. Kamaev, A survey of forecast error measures, World Applied Sciences Journal 24.24 (2013): 171-176.

[13] N. A. Zainuri, A. A. Jemain, and N. Muda, A comparison of various imputation methods for missing values in air quality data, Sains Malaysiana 44.3 (2015), 449-456.

[14] P.J. Davis, Interpolation and approximation, Dover, reprint (1975), 108–126

[15] S. J. W. Shoop, Should we ban the use of ‘Last Observation Carried forward’ analysis in epidemiological studies? SM J Public Health Epidemiol. (2015)

[16] T. Lyche, L.L. Schumaker, "On the convergence of cubic interpolating splines" A. Meir (ed.) A. Sharma (ed.) , Spline Functions and Approximation Theory , Birkhäuser (1973), 169–189

Appendix A: Data imputation graphs:

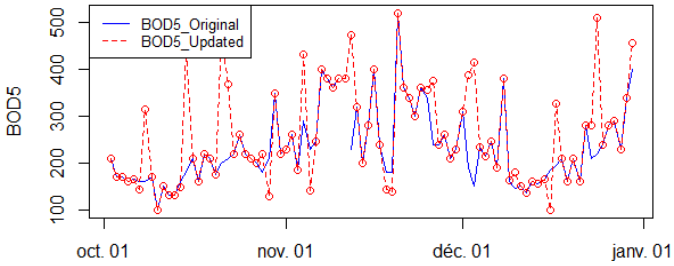


Figure A.1: Data prediction using random sample

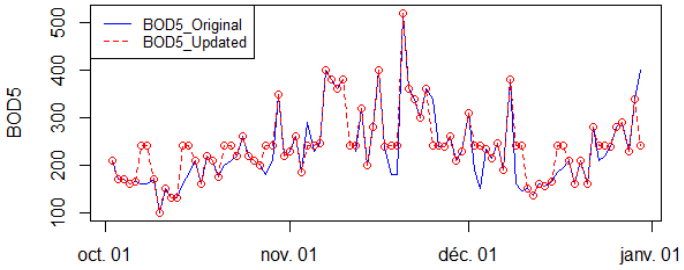
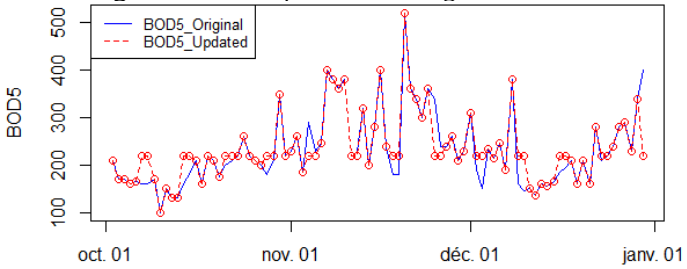


Figure A.2: Data prediction using the mean



FigureA.3: Data prediction using the median

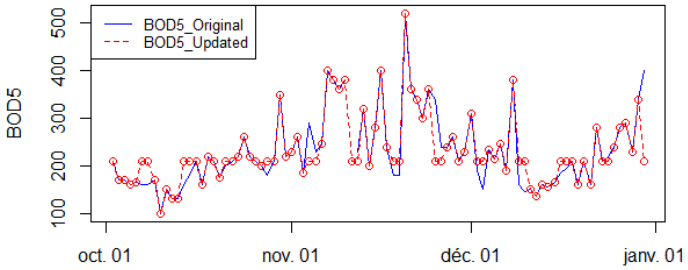


Figure A.4: Data prediction using the mode

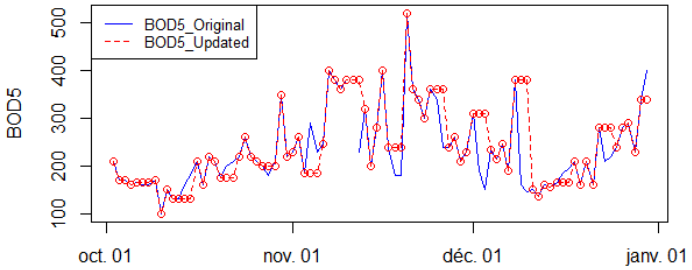


Figure A.5: Data prediction using the least observation carried forward

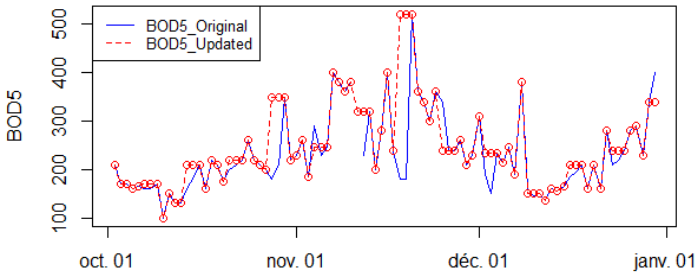


Figure A.6: Data prediction using the next observation carried backward

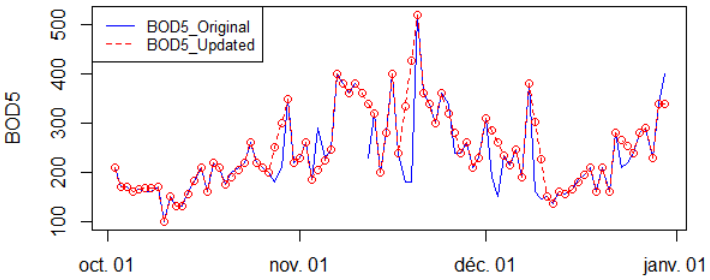


Figure A.7: Data prediction using linear interpolation

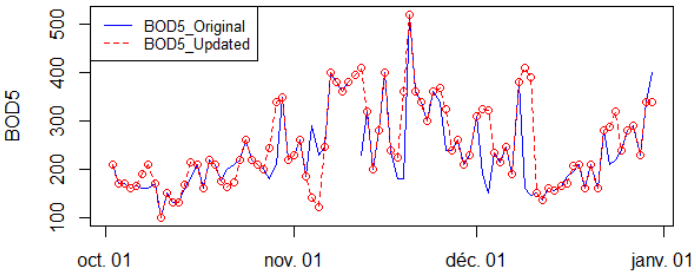


Figure A.8: Data prediction using spline interpolation

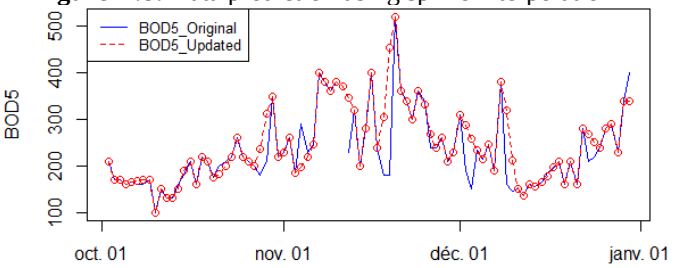


Figure A.9: Data prediction using stineman interpolation

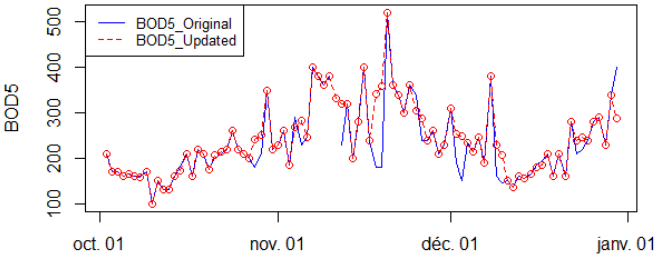


Figure A.10: Data prediction using structural model & Kalman smoothin

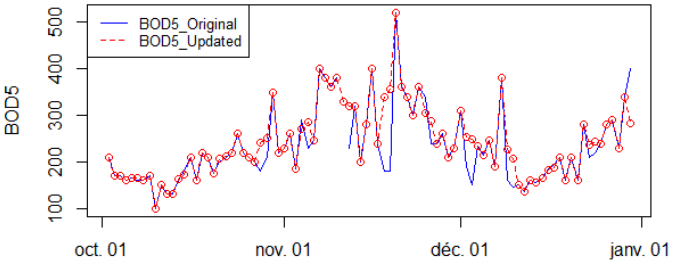


Figure A.11: Data prediction using ARIMA

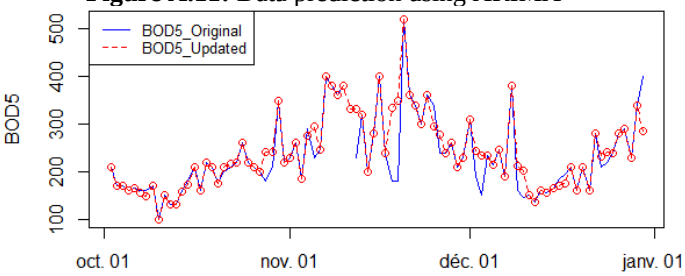


Figure A.12: Data prediction using simple moving average

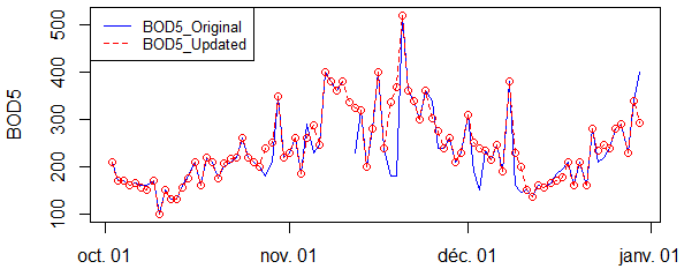


Figure A.13: Data prediction using linear weighted moving average

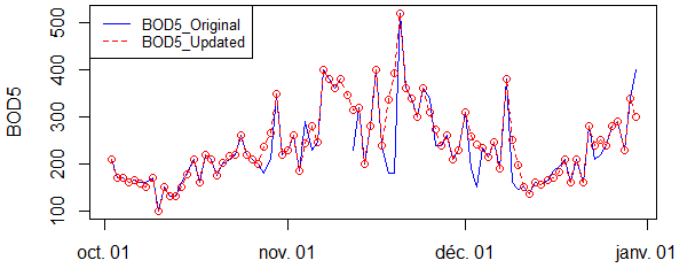


Figure A.14: Data prediction using exponential weighted moving average