poster:09

Evaluating Linear Models as a Baseline for Time Series Imputation

Rebecca Salles, Eduardo Bezerra, Jorge Soares, Eduardo Ogasawara

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

{jsoares,ebezerra}@cefet-rj.br, eogasawara@ieee.org

Abstract. *Time series prediction has been gaining attention of many researchers throughout the world for its increasing importance to preparation, planning and decision-making activities in many areas of study in science, business and government. Many data come from different sources and some of them, such as sensors, are not resilient to failures. A particular problem that occurs in these cases is the absence of data in some parts of the time series. Addressing this lack of data becomes important to enable the development of prediction models. Although there are many machine learning methods (MLM) that may be used to fill such data, there is an absence of systematically benchmarking established linear baseline methods for performance comparison. In this paper we explore linear models as baseline for time series imputation (TSI). Our results show the importance of exploring different linear approaches for TSI to encourage researchers to improve their choices for a suitable MLM for solving such problem.*

1. Introduction

Prediction is a key element to decision-making activities. Knowledge of future observations can often have a massive impact on the success or the failure of a goal. In particular, the analysis and prediction of time series attracts interest of many researchers due to its increasing importance and applications in science, business and governments.

A general strategy for making predictions based on past known values of a time series is to build a model that adequately reflects its behavior. Such model is developed based on combining data transformations and prediction methods. The latter has its parameters adjusted according to a training dataset, which is a subset of the observed time series. Once built, this model serves as a tool for predicting unknown values for that time series, including future ones.

Many data come from different sources and some of them, such as sensors, are not resilient to failures. A particular problem that occurs in these cases is the absence of data in some parts of the time series. Addressing this lack of data becomes important to enable the development of prediction models. The general problem of computing a plausible value for a missing observation in a time series to conduct an analysis with the completed data is named time series imputation (TSI) [Yozgatligil et al., 2012]. TSI can be addressed by prediction or interpolation techniques.

Although there are many machine learning methods (MLM) that may be used to fill such data, there is an absence of systematically benchmarking established baseline methods for performance comparison, particularly for non-stationary time series. There is a general concern about interpretability of MLM methods [James et al., 2013]. In this way, many scientists prefer to adopt MLM only if it clearly outperforms linear models (LM). In this paper, we evaluate LM as baseline for TSI. We have analyzed Autoregressive Integrated Moving Average (ARIMA) [Box et al., 2008], linear interpolation and spline [Zeileis and Grothendieck, 2005]. These three approaches have statistical properties that allow them to be classified as either rigid or flexible methods [James et al., 2013]. We have conducted experiments using the CATS dataset [Lendasse et al., 2004]. Our results indicate the need of exploring LM as baseline approaches for TSI to encourage researchers to improve their choices for a suitable MLM for solving such problem.

Besides this introduction, the remainder of this paper is organized as follows. Section 2 discusses fundamentals of TSI. Section 3 explores LM comparing them with state of the art MLM. Section 4 concludes the paper.

2. Time Series Imputation

A time series can be defined as a set of data of an object of interest collected over time [Box et al., 2008]. Formally, a *time series* t is a series of values $< t_1, ..., t_m >$, where $|t|$ is the number m of elements in t, and t_m is the most recent value in t. A *subsequence* r of size n in a time series t is a series of values $< v_1, ..., v_n >$, such that there exist $i_1 < i_2 < \ldots < i_n$ integers in which $v_1 = t_i,\ v_2 = t_{i+1}, \ldots,\ v_n = t_{i+n-1}$, $|r| = n$. Formally, $r = subseq(t, i, n)$.

A *gap* is a subsequence r_k in t in which all values are *NA* (Not Available). The problem of *imputation* consists in filling the set of gaps inside of a time series t with appropriate values. When the gap occurs in the last sequence of values inside a time series, the problem of imputation is exactly the same as predicting future data.

There exist a substantial variety of prediction and imputation models. We can generally consider a method for estimating a time series model to be either *rigid* or *flexible* [James et al., 2013]. Rigid methods make an initial assumption about the characteristics of the time series model; therefore the modeling process becomes a problem of estimating a set of coefficients. Nevertheless, this initially assumed model might not reflect the available time series observations [James et al., 2013]. Examples of rigid methods include linear regressions and ARIMA.

In contrast, flexible methods focus on fitting a model such that the available time series observations are approximated as much as possible taking into account a certain degree of smoothness. Some examples of flexible methods are the regression spline, linear interpolation, SVM and neural networks. In the case of prediction and imputation, although a straightforward conclusion would be that flexible are better suited than rigid methods, it is not guaranteed due to the possibility of over-fitting the time series data [James et al., 2013]. Thus, rigid methods should not be neglected when it comes to prediction/imputation.

Another very important question related to imputation and prediction is the interpretability of the methods. LM are commonly simpler to interpret than MLM. In this way, managers and researchers tend to feel safer when they understand data models. In this sense, they might prefer to adopt LM if the advantage of MLM is not statistically significant or just slightly better. Considering these characteristics, the following subsections details general LM and its implementation in R according to their rigidity level.

2.1. ARIMA

The ARIMA(p, d, q) model [Box et al., 2008] is one of the most important rigid methods for time series prediction and is derived from a composition of the autoregressive (AR) and moving average (MA) modeling processes, respectively represented by p and d, with the addition of a preliminary differentiation process (I) represented by d. Forecasting and back-forecasting are commonly used in filling gaps in univariate time series [Weerasinghe, 2010]. The selection of optimized parameters (p, d, q) for ARIMA model is not a simple task. To address such issue, commonly function *auto.arima* from the *forecast* R-package [Hyndman and Khandakar, 2008] is applied to optimize these parameters. Function *auto.arima* also identifies if the input time series is seasonal, which allows the usage of seasonal ARIMA (SARIMA) model.

2.2. Linear Interpolation

Among univariate methods for TSI, linear interpolation is one of the most popular. Despite its restrictions with regard to gap lengths [Junninen et al., 2004], its performance tend to be stable and of reasonable quality. Linear interpolation fits a straight line between the endpoints of a gap, so its equation is used to straightforwardly compute missing values [Junninen et al., 2004]. We have performed linear interpolation using the *na.approx* function of the *zoo* R-package [Zeileis and Grothendieck, 2005], which replaces NAs by linear interpolation using the function *approx* from the *stats* R-package.

2.3. Spline

In the spline interpolation method for TSI, cubic polynomials are fitted to a time series. The fitted function and its first two derivatives must be continuous at the knots, that is, where piecewise portions join [Junninen et al., 2004]. A cubic spline with knots at x_i, $i = 1, \dots, n$ is defined as $f(x) = a_i + b_i x + c_i x^2 + d_i x^3$ [Junninen et al., 2004]. The performance of spline imputation is also restricted to gap lengths, in the sense that it may present similar quality of that of a linear imputation method for short gaps, however, the performance of splines considerably decline as the length of gaps increase due to overfitting the data [James et al., 2013]. We have performed spline interpolation using the *na.spline* function of the *zoo* R-package, which replaces NAs by cubic spline interpolation using the function *spline* from the *stats* R-package.

3. Experiment Evaluation and Discussion

We have conducted an initial evaluation of linear models using the CATS Competition data set. The CATS Competition presented an artificial time series with 5,000 points, among which 100 are unknown. The competition proposed that the competitors predicted the 100 unknown values from the given time series, which are grouped into five non-consecutive gaps of 20 successive values. The CATS Competition time series is depicted in Fig. 1, in which the five gaps of unknown values (981-1000, 1981-2000, 2981-3000, 3981-4000 and 4981-5000 observations) may be observed.

The performance evaluation in the CATS Competition is based on the mean squared error (MSE) computed on the 100 unknown values (E_1) and on the 80 first unknown values (E_2). The expressions for E_1 and E_2 are presented in Equation 1, where i_b and j_b correspond to the first and last elements of the bth gap of imputed values, respectively. The second criterion (E_2) includes only the scenario in which both sides of the time

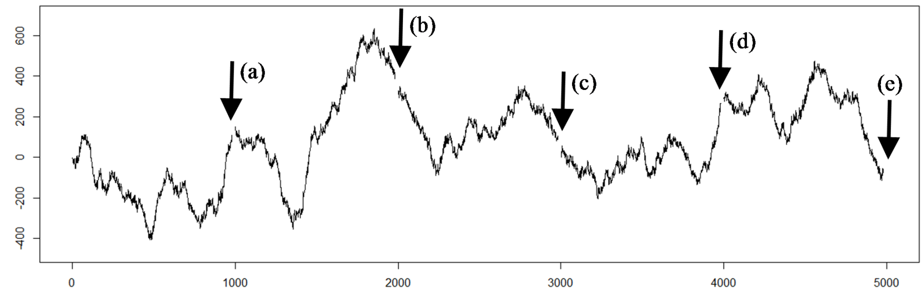

Figure 1. CATS Competition time series with five gaps

series, *i.e.*, before and after the endpoints of the gaps, can be used to perform imputation. Whereas the first criterion (E_1) includes the last gap, which presents a typical scenario of prediction. In this case not all imputation methods are capable to be used.

$$E_1 = \sum_{b=1}^{5} \left(\frac{\sum_{t=i_b}^{j_b} (x_t - \hat{x}_t)^2}{100} \right), E_2 = \sum_{b=1}^{4} \left(\frac{\sum_{t=i_b}^{j_b} (x_t - \hat{x}_t)^2}{80} \right) \quad (1)$$

Besides the previously mentioned imputation methods, we have also applied the simple method of carrying the last known time series observation forward into the gap via the *na.locf* function of the *zoo* R-package [Zeileis and Grothendieck, 2005]. Furthermore, we have included other general-purpose imputation package, namely Amelia. The method provided by the Amelia R-package implements a bootstrapping-based EM algorithm, where the means and covariance matrices of the missing data are estimated iteratively [Junninen et al., 2004]. Amelia supports the processes of imputing cross-sectional surveys, time series data, and time series cross-sectional data [Jerez et al., 2010].

The goal of adopting LM as baseline is to enhance imputation performance of MLM during training-testing, prior to its actual usage with unknown data. However, for sake of comparison purpose only, we adopt all 4900 observations as training set and the 100 unknown competition values as test set. The MSE errors for each of the 5 gaps in the CATS dataset as well as the computed values for E_1 and E_2, with respect to each applied imputation method are presented in Table 1. As expected, since the majority of imputation methods apply interpolation techniques in order to predict the first four gaps, the MSE of these gaps are generally inferior to the prediction error for the fifth block, for whose prediction we could not apply the same interpolation methodology. This is exactly the case of ARIMA. We have used the *arimainterp* function of the *TSPred* R-package [Salles and Ogasawara, 2015] to perform the imputation of the first four gaps applying forward and backward forecasting, while in the last gap, we only applied forward forecasting. Linear Interpolation was the best LM for CATS dataset, however, just like Spline, it could only be used for the first four gaps. Finally, we highlight the performance of the Amelia method, which was two orders of magnitude inferior to the one of the linear interpolation. This result is in agreement with Friese et al. [2013] who observed that Amelia works best on multidimensional time series.

The comparison between the explored LM and MLM for imputation can be observed in the ranking in Table 2, derived from the complete CATS Competition ranking [Lendasse et al., 2004]. Although ARIMA maintained its general characteristic perfor-

Table 1. MSE prediction errors for each gap with E1 and E2 measures

Method	Gap 1	Gap 2	Gap 3	Gap 4	Gap 5	E1	E2
Linear Interpolation	142	175	651	495	NA	NA	366
ARIMA	576	532	1979	583	2196	1173	917
na.locf	382	2153	3389	1103	1769	1759	1757
Spline	2014	9998	3297	939	NA	NA	4062
Amelia	51890	27480	14026	25104	24302	28560	29625

mance as a simple and rigid linear model it has obtained reasonably better results when compared to other MLM imputations. However, the linear interpolation stands out for having produced results that would lead it to the fifth position of the ranking of CATS competition.

Table 2. CATS Ranking according to E2 measure

Participant	E2	Participant	E2
1. Wichard et al.	222	11. Verdes et al.	442
2. Cellier et al.	278	12. Maldonado et al.	542
3. Särkkä et al.	346	13. Chan et al.	677
4. Simon et al.	351	14. Beliaev et al.	762
5. ***Linear Interpolation***	**366**	15. Yen-Ping et al.	894
6. Hu et al.	370	16. ***Arima (auto-arima)***	**917**
7. Palacios-González	395	17. Kong	994
8. Cai et al.	402	18. Crone et al.	995
9. Wang	402	19. Acernese et al.	1229
10. Kurogi et al.	418	20. ***na.locf***	**1757**

The characteristic one seeks on a satisfactory baseline TSI method is, firstly, the capacity to serve as parameter to define a minimum acceptable level of performance and therefore permit the evaluation of viability and expediency of TSI method. Additionally, a reliable and well-established baseline method should be able to offer a way to demonstrate and ensure the merit of methods which present high quality of performances. The value of these methods are then ratified by its better results when compared against those of baseline methods.

We can observe that our best-trained LM maintained a relevant position presenting a better performance than a reasonable number of MLM. Furthermore, the imputation errors of LM ratified the importance of the results of the best methods of the rankings, which were superior to LM. These observed properties together with its several previously discussed advantageous characteristics, such as linearity, interpretability and reliability, therefore, makes LM suitable choice for an exceptionally adequate baseline for TSI methods.

4. Conclusion

In this paper we have conducted an initial evaluation of LM as baseline for TSI, particularly when compared to MLM, as they are commonly easier to interpret. We have evaluated both rigid (ARIMA) and flexible (linear interpolation and spline) models using the CATS dataset. Although the performance of any interpolation time series method may depend greatly on the characteristics of data [Junninen et al., 2004], linear interpolation stood out as the best evaluated LM in our experiments. Since the best LM outperformed the majority of MLM that participated in the CATS competition, our results indicate the need for exploring LM as baseline for imputation during training and testing. Hence, during CATS competition, for example, should LM had been used as baseline, many poor imputations would have been avoided.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, N.J, 4 edition edition.

Friese, M., Stork, J., Guerra, R., Bartz-Beielstein, T., Thaker, S., Flasch, O., and Zaefferer, M. (2013). UniFIeD Univariate Frequency-based Imputation for Time Series Data. Technical report.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3):1–22.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer, 1st ed. 2013. corr. 4th printing 2014 edition edition.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105–115.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.

Lendasse, A., Oja, E., Simula, O., and Verleysen, M. (2004). Time Series Prediction Competition: The CATS Benchmark. In *IJCNN 2004, International Joint Conference on Neural Networks*, volume 2, pages 1615–1620, Budapest, Hungary.

Ogasawara, E., Martinez, L., de Oliveira, D., Zimbrao, G., Pappa, G., and Mattoso, M. (2010). Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Salles, R. P. and Ogasawara, E. (2015). TSPred: Functions for Baseline-Based Time Series Prediction.

Weerasinghe, S. (2010). A missing values imputation method for time series data: an efficient method to investigate the health effects of sulphur dioxide levels. *Environmetrics*, 21(2):162–172.

Yozgatligil, C., Aslan, S., Iyigun, C., and Batmaz, I. (2012). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology*, 112(1-2):143–167.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14(i06).