Time Series Forecasting for Purposes of Irrigation Management Process

Dieinison Braga[1], Ticiana L. Coelho da Silva[1], Atslands Rocha[1], Gustavo Coutinho[1], Regis P. Magalhães[1], Paulo T. Guerra[1], Jose A. F. de Macêdo[1]

[1]Federal University of Ceará (UFC)
Ceará – Brazil

dieinison@alu.ufc.br, {gustavolgcr, jose.macedo}@lia.ufc.br

{ticianalc, regismagalhaes, atslands, paulodetarso}@ufc.br

Abstract. *Irrigated agriculture is the most water-consuming sector in Brazil, representing one of the main challenges for the sustainable use of water. This study proposes and experimentally evaluates univariate time series models that predict the value of reference evapotranspiration, a metric of the water loss from crop to the environment. Reference evapotranspiration plays an essential role in irrigation management since it can be used to reduce the amount of water that will not be absorbed by the crop. The experiments performed under the meteorological dataset generated by a weather station. Moreover, the results show that the approach is a viable and lower cost solution for predicting ET_0, since only a variable needs to be monitored.*

1. Introduction

Population growth and changes in climate directly impact on worldwide food security. One of the primary objectives of agricultural research is to find improved ways to produce food. According to [Thiago et al. 2017], 72% of freshwater is consumed in irrigation, in Brazil. It is estimated that a massive portion of this amount is wasted due to poorly executed irrigation and lack of control from farmers about the exact amount of water to use in irrigation process.

Evapotranspiration value (ET_m) plays a key role in support to decision making in irrigation management, which is the simultaneous occurrence of evaporation and transpiration processes in a crop, measured in millimeters per a unit of time. We use the following equation to compute it: $ET_m = K_c \times ET_0$, where K_c is the crop coefficient c, given at INMET website[1], ET_0 is the reference crop evapotranspiration, which corresponds to the evapotranspiration rate of a grass surface. The value of ET_0 is very relevant to management and scaling in irrigation since it gives the information of how much water the crop loses to the environment [Thiago et al. 2017].

The traditional *Penman-Monteith* method [Allen et al. 1998] used to compute ET_0 is complex and does not tolerate the unavailability of some of its variables, which makes its use unfeasible. The paper [Caminha et al. 2017] proposes a Machine Learning-based approach to forecast ET_0 based on Linear Regression [James et al. 2013] and M5P [Wang and Witten 1996]. Despite the good results obtained in both techniques, they are

[1]http://sisdagro.inmet.gov.br/sisdagro/app/monitoramento/bhc

multivariate models, which means that it requires a weather station with many sensors to capture all the required variables, and there is no guarantee that models will fit, as well as in the absence of some variables.

Experiments performed by [Siami-Namini and Namin 2018] with univariate time series model demonstrated the Autoregressive Integrated Moving Average (ARIMA) [Box et al. 2015] model as a promising technique to achieve good accuracy performance in the forecast of financial time series. ARIMA model aims at describing the correlations in the data with each other. An improvement over ARIMA is Seasonal ARIMA (SARIMA) [Box et al. 2015], which takes into account the seasonality of dataset and was successfully used in short-term forecast [Tseng and Tzeng 2002]. In this paper, we use both approaches in our experiments.

The key contributions of this paper are: (i) offer an accurate and lower cost solution to estimate ET_0, since only a variable needs to be monitored; (ii) compare the performance of ARIMA, SARIMA, Linear Regression and M5P with respect to minimization achieved in the error rates in prediction; and (iii) release the dataset used in this work, for research and possible improvements by the scientific community.

The remaining sections of this article are organized as follows. Section 2 explains our proposed approach. Section 3 shows the steps necessary to accomplish our goals. Section 4 presents our experiments and its analysis. Finally, Section 5 summarizes this work and proposes future developments.

2. Time Series Forecasting

A time series (*TS*) is a series of data records indexed by dates. A time series model supposes that a series Z_t could be defined as $Z_t = T_t + S_t + \alpha_t$, being T the tendency, S the seasonality and α the white noise, at a moment t [Brockwell and Davis 2016]. Most of the *TS* models work on the assumption that the TS is stationary, i.e., its statistical properties such as mean and standard deviation remain constant over time. Due to many real-time series being non-stationary, statisticians had figured out ways to make *TS* stationary [Box et al. 2015].

In particular, differencing operator (∇) is a simple and efficient operator to transform a non-stationary TS to stationary. It is defined by the equation: $\nabla Z_t = Z_t - Z_{t-1}$, where Z is a TS at a moment t [Brockwell and Davis 2016]. In other words, we take the difference of the observation at a particular instant t with that at the previous instant $t-1$.

The ARIMA model takes three hyper-parameters p, d, q, which capture the key elements of the model, which are: (i) Autoregression (*AR*), a regression model that uses the relationship between an observation and a number (p) of lagged observations; (ii) Integrated (*I*), the number (d) of differentiation required to obtain stationarity; (iii) Moving Average (*MA*), an approach that takes into accounts the dependency between observations and the residual error terms when a moving average model is used for the lagged observations (q) [Box et al. 2015, Tseng and Tzeng 2002].

The SARIMA model incorporates both seasonal and non-seasonal factor in a TS data, its signature is $SARIMA(p, d, q) \times (P, D, Q)S$, where p and P are the non-seasonal and seasonal AR order; d and D are the non-seasonal and seasonal differencing; q and Q are the non-seasonal and seasonal MA order; and S is the time span of repeating seasonal

pattern, respectively [Tseng and Tzeng 2002].

3. Methodology

3.1. Data Collection and Cleaning

The climatic data were collected by a weather station, in the period from January, 1st to November, 29th of 2017 in the city of Quixadá, Ceará, Brazil. The original dataset contains 7941 hourly records, and it is composed of the features described in Table 1. This dataset is available in `https://github.com/Dieinison/ProjectET0/blob/master/dataset.csv`.

Table 1. Samples from dataset

Date	Atmospheric pressure		Air temperature			Relative humidity			Solar radiation		Temperature		Preciptation	Wind Speed	ET_0
	Max.	Min.	Max.	Min.	Mean	Max.	Min.	Mean	Total	Mean	Max.	Min.			
2017-11-29	620.5	599.7	21.4	19.6	32	55.2	45.3	50.1	1610	12.7	21.4	19.6	0.0	1.58	0.095
2017-11-29	620.2	599.7	21.7	19.4	32	52.3	41.9	46.9	1638	11.9	21.7	19.4	0.0	1.73	0.109
2017-11-29	620.4	599.6	20.9	19.1	34	45.8	39.7	42.3	1620	19	20.9	19.1	0.0	2.10	0.147
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

We aggregated the original hourly data on a daily basis. Furthermore, we detected outliers observations through *Proximity-Based Outlier Detection* technique [Tan et al. 2006] and remove them. The tuples contain the values described in Table 2 were removed. At the end of this procedure, 333 tuples remained.

Table 2. Removed instances

Precipitation ≥ 60
Minimum temperature ≤ 0
Minimum relative humidity ≤ 20

3.2. Prediction models

To create the prediction models, we split the dataset into 80% for training and 20% for testing. Each algorithm produced its particular model using the attributes taken as input. Thus, we generated four distinct models, Linear Regression and M5P were created from all the attributes of the dataset, ARIMA and SARIMA models were generated only with ET_0. These models and their comparisons are presented in Section 4.

For purposes of comparisons between the models generated, we used the same dataset (given by weather station from UFC Quixadá). We performed the prediction models by applying the Linear Regression and M5P algorithms, both implemented in the WEKA[2] tool.

In order to forecast through ARIMA and SARIMA, we perform the Box-Jenkins methodology [Box et al. 2015], defined as: (i) identification of the model, i.e., finding the appropriate orders for p, d, q, P, D, Q, S; (ii) estimation of the unknown parameters; (iii) validation of the model; and (iv) forecast future outcomes based on the known data.

[2]https://www.cs.waikato.ac.nz/ml/weka/

3.3. Models Evaluations

To evaluate both techniques, the Root Mean Squared Error(RMSE) and Mean Absolute Error (MAE) are calculated as the evaluation metrics of the performance, defined by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad , \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where i is the sample index, n is the total number of observations, y is the expected attribute value and $\hat{y}$ is the value output by the algorithm used [James et al. 2013]. Both metrics can range from 0 to ∞. They are negatively-oriented scores, which means lower values are better. RMSE has the benefit of penalizing large errors, while MAE is a measure of average error.

4. Experiments and Results

As stated earlier, these experiments used a real dataset with observations collected from a weather station located in Campus UFC Quixadá, in Brazil.

Initially, we generated the Machine Learning-based approaches, through WEKA tool. Due to lack of space, we do not present in this paper our Linear Regression and M5P prediction models. They are available in `http://bit.ly/result_linear_regression` and `http://bit.ly/result_m5p`, respectively.

With the view to generate time series models, we checked stationarity by plotting rolling average and rolling standard deviation as shown in Fig.1. The evaluated mean and standard deviation show significant instability over time, suggesting the data is non-stationary. Another technique to evaluate the non-stationary is the Dickey-Fuller (DF) test. The DF is a unit root test that evaluates the strength of trend in a time series component [P. Avishek 2017]. The output for DF test is shown in Table 3. As we can see, the DF Statistic is higher than the critical values, so this series is non-stationary. Therefore we can approach this with ARIMA models.

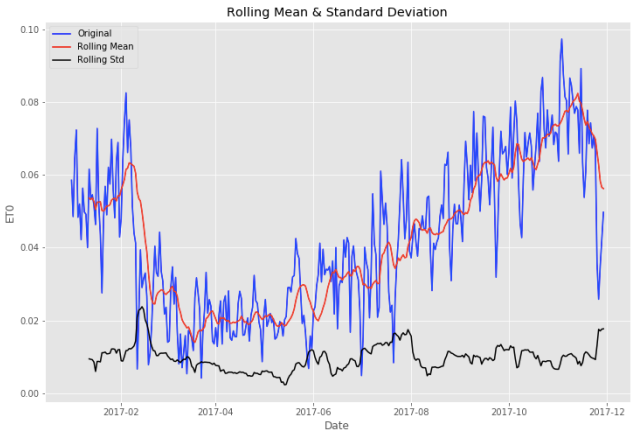

Table 3. Results of DF Test

DF Statistic	-1.695411
Critical Value 1%	-3.450695
Critical Value 5%	-2.870502
Critical Value 10%	-2.571545

Figure 1. Original ET_0.

In order to obtain the optimal hyper-parameters for ARIMA and SARIMA models, we used a function, called *auto arima*, from Pyramid[3], an API under MIT License

[3] `https://github.com/tgsmith61591/pyramid`

that provide an systematic approach to find the best hyper-parameters, based on a given information criteria, which in this case will be the Corrected Akaike Information Criterion (AIC_c), as recommended in [Brockwell and Davis 2016]. This criterion includes a penalty term to discourage the fitting of too many parameters, i.e., the fitted model with the smaller value of AIC_c will be the best choice [Smith 2017, P. Avishek 2017]. Tables 4 and 5 present the parameters output by *auto arima* function for ARIMA and SARIMA models, respectively.

Table 4. ARIMA parameters.

Parameter	Value
AR order p	1
Difference order d	1
MA order q	1

Table 5. SARIMA parameters.

Parameter	Value
AR order p	1
Difference order d	1
MA order q	1
Seasonal AR order P	0
Seasonal difference D	1
Seasonal MA order Q	2
S	12

Table 6 shows RMSEs and MAEs generated from models. As we can see, the univariate ARIMA and SARIMA models presented error values very low as it is close to zero. A value of RMSE or MAE equals to zero would that the estimator is predicting observations with perfect accuracy. Besides, in Table 7, we showed statistical properties of our label variable, ET_0, thus, as errors rates (RMSE and MAE) are less than the standard deviation, our results indeed show a good accuracy [Legates and McCabe 1999].

The results show an outperformance of multivariate model M5P, under RMSE and MAE metrics, over univariate time series models. Nevertheless, univariate time series models show us that these models indeed fit well the data, since there were small differences between predictions and expected values. Regarding *TS* models, ARIMA outperformed SARIMA in both metrics, indicating that our data is better fitted by a non-seasonal model.

Table 6. Metrics comparisons between techniques.

Model	RMSE	MAE
ARIMA	0.0196	0.0173
Linear Regression	0.0072	0.0056
M5P	0.0070	0.0056
SARIMA	0.0225	0.0201

Table 7. Mean and Standard Deviation of observed ET_0.

Statistic	Value
Mean	0.0430
Standard deviation	0.0462

Due to the costs of owning a weather station with many sensors, capture all the variables required for multivariate models might not be affordable for low-income farmers. In contrast, the results show us that an ARIMA model is an affordable solution for predicting ET_0 since only a variable needs to be monitored, with no need of multiples sensors.

5. Conclusion

This paper compares the accuracy of univariate ARIMA and SARIMA models with multivariate Machine Learning-based algorithms, Linear Regression and M5P. The results show that M5P outperform the other techniques. Despite that, this paper advocates the benefits of applying univariate time series algorithms to predict ET_0, since these models presented small differences between predictions and expected values, i.e., good accuracy. Besides, *TS* models might be an affordable solution for low-income farmers, since only a variable needs to be monitored. For future works, we aim at improving and validating our proposed models for other datasets and compare with deep learning based approaches.

Acknowledgment

The authors acknowledge FUNCAP for the research supported.

References

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., et al. (1998). Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. *FAO, Rome*, 300(9):D05109.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons, New Jersey, USA.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to time series and forecasting*. Springer, Switzerland.

Caminha, H., Silva, T., Rocha, A., and Lima, S. (2017). Estimating reference evapotranspiration using data mining prediction models and feature selection. *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017)*, 1:272–279.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer, New York, USA.

Legates, D. R. and McCabe, G. J. (1999). Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1):233–241.

P. Avishek, P. P. (2017). *Practical Time Series Analysis: Master Time Series Data Processing, Visualization, and Modeling using Python*. Packt, Birmingham, UK.

Siami-Namini, S. and Namin, A. S. (2018). Forecasting economics and financial time series: Arima vs. lstm. *arXiv preprint arXiv:1803.06386*.

Smith, T. G. (2017). *Pyramid: ARIMA estimators for Python*. MIT, USA.

Tan, P.-N. et al. (2006). *Introduction to data mining*. Pearson Education India, India.

Thiago, H. F., Wagner, M. d. C., and Marcus, A. F. (2017). *Atlas irrigação: uso da água na agricultura irrigada*. Agência Nacional de Águas, Brasília, DF, Brasil.

Tseng, F.-M. and Tzeng, G.-H. (2002). A fuzzy seasonal arima model for forecasting. *Fuzzy Sets and Systems*, 126(3):367–376.

Wang, Y. and Witten, I. H. (1996). Induction of model trees for predicting continuous classes.