Unevenly time series modelling of water flow data: a first approach

Tiago André Anciães dos Santos
tiago.a.a.santos@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2020

Abstract

Water is essential for the survival of all living organisms on Earth, and hence it is of the utmost importance to preserve it and do a sustainable management of the hydric resources at our disposable. The main goal of the WISDom (Water Intelligence System Data) project is to create models and algorithms for data analysis in order to reduce water losses and improve the supply of drinking water, which will assist in the decision making process by management entities. Within the scope of this project, this paper has as one of its main objectives to model a flow series provided by Infraquinta (one of the managing entities partner of this project) and to forecast the next observations, using the series provided and maintaining one major characteristic, its irregularity. This modeling was done by six methods: three in which the equally spaced series obtained by aggregating the original series (most common method) was used and three considering the series with different spacing. Another major objective is the comparison of these two modeling approaches, by considering the quality of the predictions, its complexity and time spent in obtaining the models. Results show that there is little to no advantage in using this alternative approach. As for the forecasts, although better in some cases, were similar for both methods and these "new" methods were proven to be more complex and more expensive computationally, even though the difference is small in this last field.
Keywords: Irregular Time Series, Water flow data, Forecast, Modified Holt method, Modified Holt-Winters method

1. Introduction

One of the goals of this analysis is to develop prediction models for the water flow using different methods. We will use two techniques: the most common is to convert the time series into a regular and then model it using some well-known methods, the other technique, and the focus of this paper, is to model directly the irregular time series without changing the spacing between observations. In order to make predictions for the regular case, we will consider Seasonal Autoregressive Integrated Moving Average (SARIMA), Holt's method and Holt-Winters' method, all three are widely used when it comes to model time series. For the irregular time series we will use Singular Spectrum Analysis (SSA) and two other models also used for the equally spaced time series (Holt and Holt-Winters' methods) but this time with some modifications that allow us to use them with unevenly time series.

With this done, our main goal is to compare the results obtained by the two different techniques mentioned before and see if there is any advantage in using this less traditional technique regarding the error of the predictions and the time needed to produce these forecasts using this less traditional technique.

Here we analyze different time series and this way give different inputs to our methods. First we do a simulation study where we generate several regular time series with different parameters. This helps us to see if the new methods are leading to good results before entering in more complex time series.

The second set of time series that we use has two monthly series, one regarding the level of water at Lake Erie and one regarding the average temperatures in New York City. Both of them are equally spaced and use them like they are for the three models that can only deal with this type of series. For the other three methods we modify the series turning them into unevenly time series and this is the first real test to these methods. In this paper we will not show those results.

The last time series that we analyze, and the one that this study is focus on, was provided by Infraquinta and it contains water flow observations. This series is unevenly, so this is the only series in

this paper that is used to perform the two techniques referred earlier in this text. Hence, we will give a special attention to the results that come from the analysis of this last series.

2. State-of-the-art

Researchers have been providing us with several methods to model unevenly time series in the past years, Cipra and Hanzák are the biggest names linked to those achievements but in recent years. As pioneers we have names as Wright that was involved in the Simple exponential smoothing (SES) and Holt method for irregular time series [12], Aldrin and Damsleth works in SES, Holt method and DES [1]. Ratinger also made his contribution with Holt-Winters method for the case of a single gap in observations [10]. Cipra alone and in collaboration with other researchers including Hanzák develop Holt-Winters method for time series with missing observations [11], Holt method with exponential or damped linear trend for irregular time series [3], Double exponential smoothing (DES) for irregular time series [3], exponential smoothing of order m for irregular time series [8], SES for irregular time series [8] and exponential smoothing in L1 norm and M-estimation [2].

Recently Andreas Eckner worked in some tools that helps as in the analysis and manipulation of the irregular time series [4]. Some of his important achievements are the trend and seasonality estimation [5] and some algorithms such as: moving averages and other rolling operators [6].

3. Background

4. Unevenly Time Series

In this subsection we will talk about a specific type of time series, unevenly ones, these have values collected in a discrete way and the collecting times are not equally spaced (sometimes the difference between two consecutive times can be equal but it is not the rule). From now on we will denote an unevenly time series X by $\{(t_n, X_n) : 1 \leq n \leq N(X)\}$ or $\{X_{t_n} : 1 \leq n \leq N(X)$ [1]$\}$ where [4]:

- $T(X) = (t_1, ..., t_{N(X)})$ is the set with strictly-increasing observation times;
- $(X_1, ..., X_{N(X)})$ are the observation values;
- $N(X)$ is the length of the time series.

For the sampled value of X at time t we can define some quantities:

- $X[t]_{last}$ is the last observation value of X at or before time t;
- $X[t]_{next}$ is the next observation value of X at or after time t;
- $X[t]_{linear}$ is the linear interpolated value of X at time t.

Bearing this in mind, we can easily say that when we are considering some time $t \in T(X)$ we have that $X[t]_{last} = X[t]_{next} = X[t]_{linear} = X_t$.

In time series analysis the majority of methods can only handle with evenly-spaced time series and it is really not so straightforward to extend to unevenly data. One common approach is to convert irregular time series data into regular by adding missing values and then estimate or interpolate those missing values. Other approach is to aggregate the irregular time series in equally spaced time stamps, this will be the approach used in this work. Another approach, but more complicate due to the difficulty of estimating, is to apply continuous time series models [9].

The following provides a simple description of methods to obtain a computationally simple solution for extract usefulness information of unevenly time series.

4.1. Rolling Time Series Operators

These operators allow us to extract some information about the time series that we are working on for a certain time window inside our period interval. In each step, we can extract some statistics of the data such as the sum/average of observations' values, the number of observations inside that window, among others. Then we update the window by shifting it to the right and maintaining the length of the window [6].

Generic Algorithm

The first step is to choose a length ($\tau \geq 0$) for the window and a time $t \in T(X)$. Here we will consider intervals of the form $]t - \tau, t]$ where $t \in T(X)$.

The second step is to construct the window and we can define two variables: "first" which will be the index of the left-most observation inside our window and the "last" which will represent the index of the observation that happened at time t.

Having this we can build two lists, one with the observations' values using the array $V(X)$ and choosing the indexes from "first" to "last" and other with the observations' times using of $T(X)$ with the same criteria as before.

We are now in conditions to collect the information that we talked before and for the calculation of each statistics there is a different algorithm.

When we finish our analysis on that window we change our window updating the initial t to $t + 1$ and repeat all the other steps. This process should continue till t is equal to the last entry of $T(X)$.

4.2. Simple Moving Averages

Simple Moving Average (SMAs) is another algorithm to summarize the average of a given period

[1] $N(X)$ is fixed and it does not depend on the time series evolution.

of time. Like in the previous algorithm we will consider a length τ for the window and for some $t \in T(X)$ we have:

(i) $SMA_{last}(X,\tau)_t = \frac{1}{\tau}\int_0^\tau X[t-s]_{last}ds$

(ii) $SMA_{next}(X,\tau)_t = \frac{1}{\tau}\int_0^\tau X[t-s]_{next}ds$,

(iii) $SMA_{linear}(X,\tau)_t = \frac{1}{\tau}\int_0^\tau X[t-s]_{linear}ds$.

So we can calculate SMA by three different ways depending on the quantity from *Definition 4* that we choose to integrate.

The major difference between the rolling average and this new operator is that in the first case every observation had the same weight on the average calculation and here each data point has a weight depending on the spacing of observations' time. It is easily to see that if we are dealing with equally spaced time series the two operators will lead us to the same result [6].

4.3. Exponential Moving Averages

Similarly to SMAs, Exponential Moving Average (EMA) gives weight to the observations but this time more weight is given to the most recent ones. Once again we have three ways to perform its calculation [6]:

(i) $EMA_{last}(X,\tau)_t = \frac{1}{\tau}\int_0^\infty X[t-s]_{last}e^{-\frac{s}{\tau}}ds$

(ii) $EMA_{next}(X,\tau)_t = \frac{1}{\tau}\int_0^\infty X[t-s]_{next}e^{-\frac{s}{\tau}}ds$

(iii) $EMA_{linear}(X,\tau)_t = \frac{1}{\tau}\int_0^\infty X[t-s]_{linear}e^{-\frac{s}{\tau}}ds$

4.4. Non-Causal Rolling Operators

In this section we will considered a different type of interval from the one consider in the operators so far. Here we need to choose two constants τ and η to produce an interval of the form $]t-\tau, t+\eta]$ where $\tau \geq 0$ and $\eta \geq 0$, this means that we will have a two-sided rolling time window.

With this new interval we have not only past observations but also future information which can be helpful for some applications like smoothing of noisy data.

For the computation of this algorithm we will need to calculate the "first" as we have done so far and the "last", but this time with a small difference because our right extreme of the interval is no longer $t \in T(X)$ so we need to do something similar to what we do for "first" but for the right hand side of the time period.

The rest of the algorithm is similar to the rolling average one and with a small effort it is possible to change the algorithms for each statistic in order to give us its values for a two-sided window [6].

4.5. Holt method

The traditional Holt method can be extended for the irregular time series case [7]. Here we present the new updating formulas for slope and trend:

$$L_{t_{n+1}} = (1-\alpha_{t_{n+1}})[L_{t_n} + (t_{n+1}-t_n)T_{t_n}] + \\ +\alpha_{t_{n+1}}x_{t_{n+1}} \quad (1a)$$

$$T_{t_{n+1}} = (1-\gamma_{t_{n+1}})T_{t_n} + \gamma_{t_{n+1}}\frac{L_{t_{n+1}} - L_{t_n}}{t_{n+1}-t_n} \quad (1b)$$

where the coefficients α_{t_n} and γ_{t_n} are updated at each step since the space between observations rarely stays the same. The updating formulas are given by:

$$\alpha_{t_{n+1}} = \frac{\alpha_{t_n}}{\alpha_{t_n} + (1-\alpha)^{t_{n+1}-t_n}}, \\ \gamma_{t_{n+1}} = \frac{\gamma_{t_n}}{\gamma_{t_n} + (1-\gamma)^{t_{n+1}-t_n}}. \quad (2)$$

where $\alpha, \gamma \in (0,1)$ are chosen in the beginning of the computation.

Initial values

The constants α, γ belong to the interval $(0,1)$ as said before and their initial values can be optimized by trying different combinations of these two values and choosing the combination that produced better results.

In order to initialize the slope and trend values L_0, T_0, a linear regression can be fitted using a few observations from the beginning of the time series giving us these two values.

4.6. Holt-Winters method

The classic Holt-Winters method can also be changed in order to be able to deal with irregular time series [7]. So, let us consider an irregular time series $\{X_{t_n}, n \in \mathbb{Z}\}$ with linear trend and seasonal components. The equations for the forecast $\hat{x}_{t_n+\tau}(t_n)$ and smoothed values $\hat{x}_{t_n}$ are the following:

$$\hat{x}_{t_n+\tau}(t_n) = L_{t_n} + \tau T_{t_n} + S_{t_n}(t_n+\tau) \quad (3a)$$

$$\hat{x}_{t_n} = L_{t_n} + S_{t_n}(t_n) \quad (3b)$$

where L_{t_n} represents the level, T_{t_n} is the slope and S_{t_n} is the seasonal component.

In order to model the seasonal component, $K \geq 1$, different real-valued functions $f_1, f_2, ..., f_K$, all defined in $\mathbb{R}$, will be considered. It is supposed that these f_k's are periodic and each one of them has its own period $p_k \in (0,\infty)$. S_{t_n} will be a linear combination of the previous functions and it is given by:

$$S_{t_n}(t) = \sum_{k=1}^{K} A_{t_n}^k f_k(t) \quad (4)$$

where $A_{t_n}^k \in \mathbb{R}$ are the proper amplitudes for each function at time t_n.

In each time step the updating formulas for the level, slope and amplitudes are the following:

$$L_{t_{n+1}} = L_{t_n} + (t_{n+1} - t_n)T_{t_n} + \alpha t_{n+1} e_{t_{n+1}} \quad (5a)$$

$$T_{t_{n+1}} = T_{t_n} + \frac{\gamma_{t_{n+1}} \alpha_{t_{n+1}} e_{t_{n+1}}}{(t_{n+1} - t_n)} \quad (5b)$$

$$A^k_{t_{n+1}} = A^k_{t_n} + \frac{\delta^k_{t_{n+1}}(1 - \alpha_{t_{n+1}})e_{t_{n+1}}}{f_k(t_{n+1})} \quad (5c)$$

where $e_{t_{n+1}} = x_{t_{n+1}} - \hat{x}_{t_{n+1}}(t_n)$ and we take $\frac{0}{0} = 0$ in the amplitude's updating formula. On the previous formulas there are three smoothing coefficients $\alpha_{t_n}, \gamma_{t_n}, \delta^k_{t_n} \in (0, 1)$ and all of them are updated recursively over time.

For α_{t_n} updating formula, it was used the idea of exponential weighting:

$$\alpha_{t_{n+1}} = \frac{\alpha_{t_n}}{\alpha_{t_n} + (1 - \alpha)^{t_{n+1} - t_n}} \quad (6)$$

where $\alpha \in (0, 1)$ is a smoothing constant for level that is given in the beginning of the computation of the model.

For γ_{t_n} updating formula, we have the form:

$$\gamma_{t_{n+1}} = \frac{\gamma_{t_n}}{\gamma_{t_n} + \frac{t_n - t_{n-1}}{t_{n+1} - t_n}(1 - \gamma)^{t_{n+1} - t_n}} \quad (7)$$

where $\gamma \in (0, 1)$ is a smoothing constant for slope that is given in the beginning of the computation of the model. The idea of the previous formula is to avoid the impact of the time distance being close to zero in the estimation of the slope.

The last smoothing coefficient, $\delta^k_{t_n}$, has a more complex updating formula. For $k = 1, ..., K$ let us denote

$$W^k_{t_n} \equiv \sum_{j=0}^{\infty} (1 - \delta^k)^{t_n - t_{n-j}} f^2_k(t_{n-j}) \quad (8)$$

This formula can be transformed in a recursive one as

$$W^k_{t_{n+1}} = (1 - \delta^k)^{t_{n+1} - t_n} W^k_{t_n} + f^2_k(t_{n+1}) \quad (9)$$

where δ^k are smoothing constants for each f_k, in a special case $\delta^k \equiv \delta$, this constant is also defined in the beginning of the computation.

$$\Delta^k_{t_{n+1}} \equiv \frac{f^2_k(t_{n+1})}{W^k_{t^k_{n+1}}} \quad (10a)$$

$$\Delta_{t_{n+1}} \equiv 1 - \prod_{k=1}^{K} (1 - \Delta^k_{t_{n+1}}) \in [0, 1] \quad (10b)$$

$$D_{t_{n+1}} \equiv \sum_{k=1}^{K} \Delta^k_{t_{n+1}} \geq 0 \quad (10c)$$

$$\delta^k_{t_{n+1}} \equiv \frac{\Delta_{t_{n+1}}}{D_{t_{n+1}}} \Delta^k_{t_{n+1}} \in [0, 1], k = 1, ..., K \quad (10d)$$

Now for the case of a multiplicative seasonality, we have to change the prediction, smoothing and amplitude's updating formulas to:

$$\hat{x}_{t_n+\tau}(t_n) = (L_{t_n} + \tau T_{t_n}) exp[S_{t_n}(t_n + \tau)] \quad (11a)$$

$$\hat{x}_{t_n} = L_{t_n} exp[S_{t_n}(t_n)] \quad (11b)$$

$$A^k_{t_{n+1}} = A^k_{t_n} + \delta^k_{t_{n+1}}(1 - \alpha_{t_{n+1}})[ln(x_{t_{n+1}}) - ln(\hat{x}_{t_{n+1}}(t_n))] \quad (11c)$$

Initial values of slope, trend and constants

As mentioned before, the constants α, γ and $\delta^k \equiv \delta$ belong to $(0, 1]$, their initial values can be optimized trying different combinations of these three values and choosing the combination that produced better results.

The slope L_0 and trend T_0 can be initialized in the same way as in the Holt method for irregular time series talked before.

For the seasonal amplitudes the initialization will be $A^k_0 = 0$ for $k = 1, ..., K$ and for W^k_0 the following approximation will be considered:

$$W^k_0 \approx \sum_{j=0}^{\infty} (1 - \delta^k)^{jq} \overline{f^2_k} = \frac{\overline{f^2_k}}{1 - (1 - \delta^k)^q} \quad (12)$$

where $\overline{f^2_k}$ is the average squared value of f_k for all the observation times available.

Seasonality modeling function f_k

There are several functions that can be chosen to perform this task, but this choice depends on the seasonal pattern. The first choice that needs to be made is the number K. With an higher K, it is expected that we get better results even with more complicated seasonal patterns, but we have to be cautious to avoid over-fitting. Here we will model the time series with different numbers of K, to see its influence on the final model.

In this work we only used trigonometric functions of time. Here we will consider different periods depending on the number of harmonics that better represent the series. There will be a sine and a cosine function for each harmonic and both with the same period. The number of harmonics to be included will be $h = K/2$, so in this case we can only consider even K values. In order to prevent over-fitting we need to take in consideration the inequality $2h \leq p/q$. The periods of each harmonic will be taken as $p, p/2, p/3, ...$ where p is the period length of the series. For example, if we consider $K = 4$ we will have the following f_k:

$$f_1(t) = \sin\frac{2\pi t}{p}, f_2(t) = \cos\frac{2\pi t}{p}, \quad f_3(t) = \sin\frac{4\pi t}{p}, f_4(t) = \cos\frac{4\pi t}{p}. \quad (13)$$

5. Results

5.1. Simulation study

In this section, we will test all the methods used in this work. With that goal in mind, we simulated three different time series with different period length ($p = 7, 12$ and 24).

The model used to generate the series was the following:

$$x_t = L_t + S_t + \epsilon_t, \epsilon_t \overset{iid}{\sim} N(0,1) \tag{14a}$$

$$L_t = L_{t-1} + \mu_t, \mu_t \overset{iid}{\sim} N(0, 0.1^2) \tag{14b}$$

with ϵ_t and μ_t being mutually independent. The seasonal component S_t is given by:

$$\begin{aligned} S_t = (1-\nu)(S_{t-1} + S_{t-p} - S_{t-p-1}) - \\ -\nu \sum_{j=t-p+1}^{t-1} S_j + \pi_t, \pi_t \overset{iid}{\sim} N(0,1) \end{aligned} \tag{15}$$

where $\nu \in [0,1]$ is a constant that produces a normalization of S leading it to sum up zero and a lower value of ν creates a smoother pattern. The π_t are independent of μ_t and ϵ_t. The initial values chosen were $L_0 = 0$ and $S_j = 0$ for $j = -p, ..., 0$.

The model presented for S_t is a special AR$(p+1)$ process and it is trying to do a better representation of reality. Typically in a SARIMA model the seasonal component follows a random walk for each calendar unit (the whole S_t follows an AR(p) process). This means that the seasonal indices are independent for different calendar units and the seasonal pattern is not autocorrelated or smooth, which rarely happens in reality.

For each value of p, we will simulate twenty one complete periods, so each time series will have length of $21p$. The initial values of these series will have a heavy impact of initialization of S_t, so we will not use the first $10p$ values. The following ten periods will be used as training dataset and we will test our methods in the last complete period, so we will forecast p values and calculate the RMSE value.

When generating the time series, three different values for ν were considered: $(0.05, 0.1, 0.2)$, and for each combination of p and ν one hundred time series were simulated. Naturally, to apply modified Holt, modified Holt-Winters and SSA methods we had to transform the regular simulated time series into irregular ones by taking off some of the observations.

The results, in general, were better using Holt-Winters. But we did not see much difference between these methods.

In terms of the computation times of the methods, the models where the parameters had to be "optimized" by hand took a bit longer than the ones that had a function that optimizes the smoothing constants by itself. But in general there was not much difference among them. This fact can be explained by the size of the time series used in this simulation study. For larger datasets this might be different.

5.2. Infraquinta's dataset

The series that we use here was provided by Infraquinta, which is a portuguese water utility, and corresponds to a year of observations of water flow collected by a specific sensor in the pipeline. From now on we will call this series by INFRA.

INFRA has 401.971 observations, from the year of 2017. With a brief analysis of the data, we noticed that the last day of every month was missing and after a conversation with the entity responsible of collecting the data we understood why this happened. The series was updated month by month and when they extract a month from the system, for example January, they choose the ending data as "2017-01-31 00:00:00" which did not include the last day, they should have done "2017-01-31 23:59:59" to avoid this. The solution of this problem is presented in front.

An important thing when working with irregular time series is to understand how the spacing between observations is changing during time.

The observations are mostly spaced with values in the interval $[1, 2]$ minutes. The maximum space is six minutes but there are just a few of them that reach that value. This will be helpful to solve the problem of the last day being missing.

Checking to the values of INFRA we noticed that 54 of them were negative, which cannot happen in a water flow, so we removed those observations from the dataset, leading to a total of 401.917 data points.

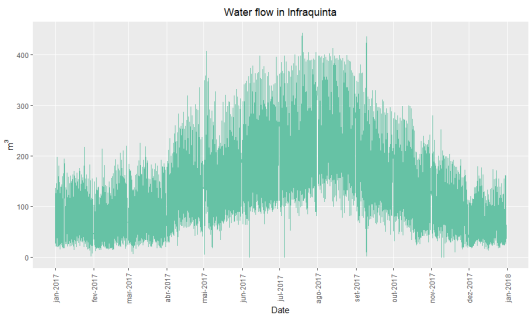

Figure 1: INFRA time series

In order to explore seasonality in the data, we made boxplots dividing the data into the different months. The results are shown in figure 2. As expected, since we are talking about water consumption, the summer months have higher values when compared to the other seasons of the year.

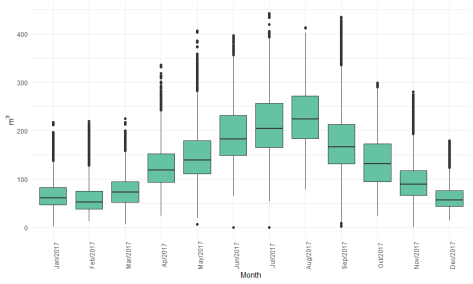

Figure 2: Boxplots by month

It is also important to check if we have a difference between the week days and the weekend. These two groups might have different behaviors because most of us change habits from the week to the weekends. In figure 3, we have plots of each week day individually and during day hours we have a lower water flow in the weekend, but in the night hours the plots are really close to each other which makes sense because most of the people is sleeping by that time and have the same habits despite the day of the week.

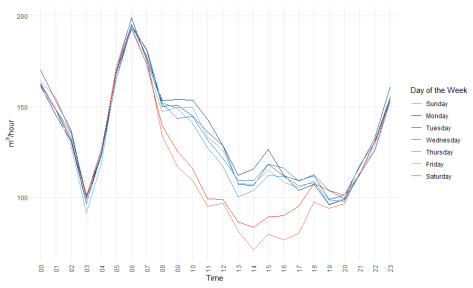

Figure 3: Water flow by day of the week

We can also check how the series behaves in respect to the hours of the day to see if we have a daily seasonality. The plot in figure 4, shows that the higher consumption happens at 6 a.m. (this was also shown in figure 3). Usually this would be a strange fact, not many people are awaken by that time, but in Infraquinta the garden irrigation system is already working at this hour, which explains this event.

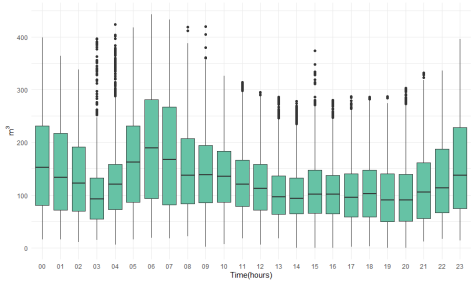

Figure 4: Boxplots by hour

Last day of the month Imputation

Given that our observations are spaced with values in the interval $[1, 2]$ minutes, we decided to impute values with one minute between the new data points, meaning that the last day of each month represents a regular time series, but this will not affect the nature of the series since the rest of INFRA will remain unevenly. First, we create the new timestamps and for each time stamp, we impute an observation based in two approaches, for each month we choose the most accurate one.

One of the approaches will take into consideration the four previous days that have the same position in the week as the last lay of that specific month. For example, the last day of January in 2017 was Tuesday, so we picked the four Tuesdays before. In the other approach we chose the two previous days with the same position in the week and the two following ones.

The next step is the same in both of them. Let's say that we want to impute the value where the timestamp has the hour "00:01:00". We will search in the four days that we have chosen before if there are any observations between the hours "00:00:30" and "00:01:30" in those days. The imputed value will be the mean of those observations that were in that interval. In case that there is not observations in those days, "NA" will be attached to that timestamp and in the end of the process "NA"s will be replaced using interpolation.

In order to choose the best method to execute this task for each month, we applied both methods to the penultimate day of the month and then compared the results to the real ones using RMSE. The method that had the lower RMSE's value was the one used to generate observations for the last day of that month. Notice that for December we could only use the first approach since we did not have information about January of 2018.

5.3. Modelling the time series

Here we will just presents the results obtained by the three methods that model irregular time series: modified Holt, modified Holt-Winters and SSA. In subsection 5.4, we will show the results for all the methods that we used in this work.

5.3.1 Modelling with Holt

In order to model the INFRA irregular series with Holt's method we will, once again, use the function that with built in R as we did for TEMP and LEVEL datasets. For the constants' "optimization" we started with $\alpha = 0.1$ and $\gamma = 0.1$ as the first pair and then tried all the combinations of these two smoothing parameters with decimal shifts. Since the best values were obtained when $\gamma = 0.1$, we decided to try the values 0.01 till 0.09 with shifts of

0.01. The combination that produced the smaller value of RMSE was $\alpha = 0.1$ and $\gamma = 0.05$.

Having the residuals we decided to check its normality. Both histogram and QQ-plot showed some violations of normality.

In figure 5 we have the fitted values in red and the real observations in blue, this produce a RMSE of 10.752. The forecast of thirty six observations showed in 6 had a RMSE of 44.174.

5.3.2 Modelling with Holt-Winters

In order to fit a Holt-Winters model for irregular time series, we need to specify some parameters such as the seasonal functions, its period and the number of harmonics (related to the value of K) that best suit the data. There are also three smoothing constants (α, γ, δ) that need to be optimized.

The seasonal functions selected to use here were the trigonometric with period $p = 1440$. To initialize and see how the method behaves with chose $K = 2$. This value was used in the constants' optimization phase, but it was changed afterwards to see if better results could be achieved.

The optimization of the three smoothing constants was not the usual one and for sure not the best one and consequently, not fully optimized. For this task, we tried different combinations of these three values starting them at 0.1 and doing increments of 0.1 till the maximum of 0.9. The other features were not changed during these computations and the results were compared using RMSE's value for the train and test set.

The best results were obtained using $\gamma = 0.1$. As the values of γ went up the RMSE went up as well to orders of 10^{20}, and with this information we decided to use $\gamma = 0.05$ producing even better results. In relation of the other two constant, it was easy to verify that for higher values of (α, δ) the best results were achieved, so in our final computation we used 0.9 for both of them.

The histogram of the residuals of the model, it has a left-skewed form with mode near zero. QQ-plot showed that residuals have heavy tails.In the plot with residuals against the fitted values we saw that when the fitted values are lower, the residuals are all spread out around zero but when the fitted values start growing the residuals get higher as well and above zero.

In the figure 7, the last 100 observations of the train set are shown, where in blue we have the real observations and in red we have the fitted values produced from the Holt-Winters method. These results correspond to a RMSE= 1.494.

The forecast showed in figure 8 had a RMSE of 19.258. We can see that the model predicted the

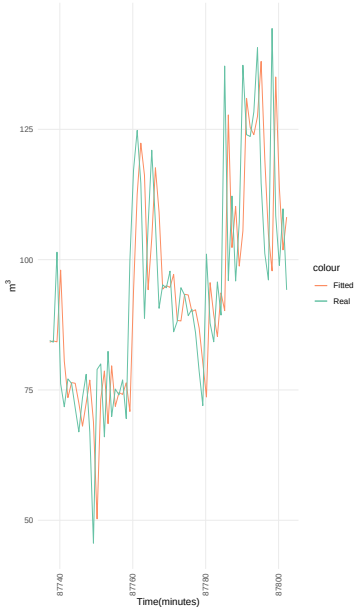

Figure 5: Fitted values of Holt for irregular series

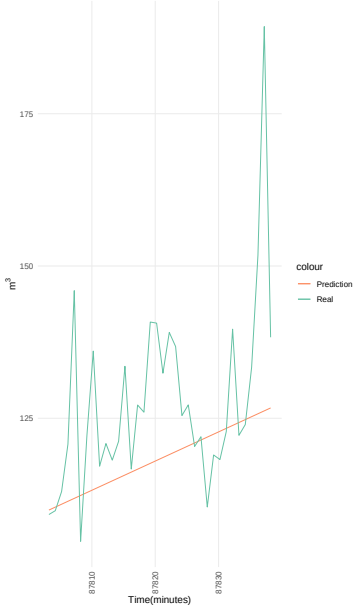

Figure 6: Forecast using Holt for irregular series

Figure 7: Fitted values of Holt-Winters for irregular series

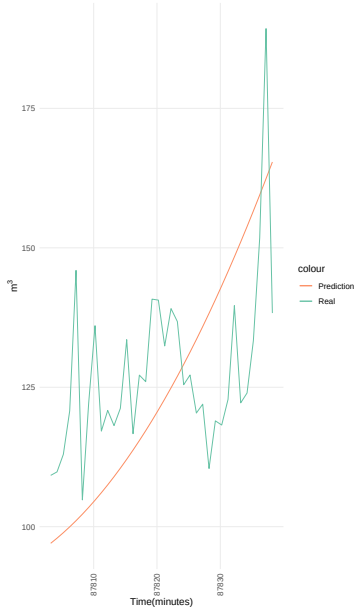

Figure 8: Forecast using Holt-Winters for irregular series

growing of the water flow but without showing the spikes that the real values show.

5.3.3 Modelling using SSA

This irregular series was also model by SSA, using the same methodology that we used when we analysed TEMP and LEVEL series.

The histogram of the residuals of the model has a right-skewed form with mode near zero. QQ-plot showed that residuals have heavy tails similar to the ones of the Holt-Winters method for irregular series. The plot with fitted values against residuals showed that residuals were all spread out around zero, but with more residuals above this value.

In terms of the fitted values showed in 9 we have the worst model comparing to the ones used here, with a RMSE of 24.479. The forecast that we see in figure 10 is similar to the one provided by both Holt's methods, but this time all the predictions were below the real values. The forecast had a RMSE of 26.811.

5.4. Comparison of the results

Here we will compare the results obtained by the methods in terms of the RMSE of the forecast and the computation time. In tables 1 and 2 we have the RMSE's values of the forecast for all methods for an easier comparation.

	SARIMA	Holt	HW
INFRA	55.048	30.141	79.693

Table 1: RMSE for the different methods for regular INFRA forecasts

	Irreg. Holt	Irreg. HW	SSA
INFRA	44.174	19.258	26.811

Table 2: RMSE for the different methods for irregular INFRA forecasts

Modified Holt-Winters provided us the best results for INFRA time series forecast, despite having the worst RMSE in the fitted values. Of course we got different errors' values among the methods, but as we can see all of them are in the same order of magnitude, so there were not a big difference between them. Looking to the forecast plots showed in this section we would probably choose Holt-Winters for irregular series as the one to forecast this series with a short length of predictions.

This time we saw a difference in terms of the time expended in the model's creation. Methods like SSA and Holt for regular time series were the fastest ones by far. This might be explain by the fact that the methods optimized the smoothing constants (in Holt's case) and created the groups (SSA's case) by

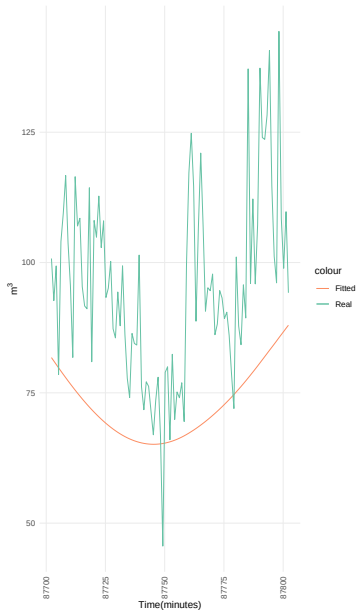

Figure 9: Fitted values of SSA

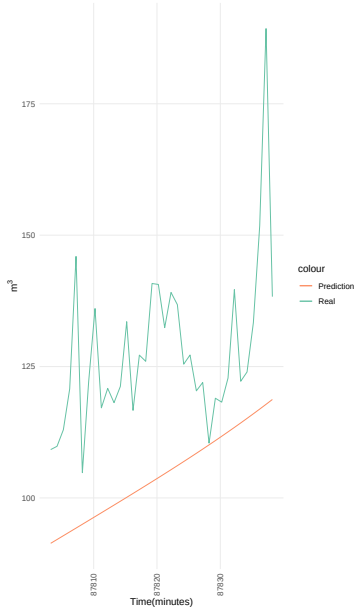

Figure 10: Forecast using SSA

themselves. In exception of SARIMA, we needed to "optimize" the parameters by hand which expended a lot of time (some hours in the case of Holt and Holt-Winters for irregular series) and some minutes for normal Holt-Winters. Although SARIMA had the function *auto.arima* to choose the best model, it also took some hours (similar to the modified Holt-Winters) to find that model because it tests several models by itself in order to choose the best one. If we already knew all the parameters in advance all the methods would model the series in less than a minute, modified Holt-Winters took a bit longer but nothing special.

6. Conclusions

6.1. Achievements

The primary objective of this work is to model and predict an irregular time series of water flow from a water utility: Infraquinta. As this time series is irregular and fewer methods exists specifically for the analysis of this type of data, we carried out a literature review and we have implemented in R two of those methods: Holt and Holt-Winters. Before we start working with Infraquinta's series, we decided to study the performance of the methods on simulated time series. To do that we simulated nine different time series (all regular) and then we model them using six different methods: SARIMA, Holt, Holt-Winters, Holt for irregular time series, Holt-Winters for irregular time series and SSA. Naturally, to apply the latest three methods we had to transform the regular simulated time series into irregular ones. A second analysis was made using two real time series, both regular, but in order to apply-Holt and Holt-Winters for irregular time series we have done as before and we took off some observations resulting in unevenly series. After this two studies we started to analyze Infraquinta's series. We had observations from a whole year (2017), but we decided to simplify our analysis by using only two months of observations of that year. Before starting the time series analysis it was necessary to clean up the data.

Concerning the use of methods for regular time series, we needed to convert our irregular series into a regular ones and this was done using aggregation. All the methods were then used to produce the forecast. The best model (regarding RMSE) was provided by Holt-Winters for irregular time series. Despite having the best result, this method could achieve an even lower value of RMSE because its parameters were not fully optimized producing that way a poorer result than what it could had been. As mencioned in subsection 5.4 the computational time needed to model the time series was way higher for the methods were we needed to "optimize" the parameters by ourselves, which was expected.

Overall, we obtained "good" results using this ap-

proach of modelling directly the irregular time series, but not much different than the ones obtained by the traditional technique of turning the series into a regular one and then apply the usual methods. Of course when we transform a time series into a equally spaced data we lost information and can introduce bias leading to incorrect predictions, but the methods are more complex and computationally costly. Holt and Holt-Winters are simple and popular forecasting procedures for regular time series but it is well known that is particularly suitable for producing short-term forecast. We saw the same behaviour of the adapted methods for irregular time series. Singular spectrum analysis is another alternative to deal with unevenly time series, but to improve its forecast ability we should consider a more deep analysis to understand the structure of the original time series. Some attempts of predicting observations with a long spacing in time showed us that the model resulted in a poorly forecast.

6.2. Future Work

In terms of future work, it would be interesting to apply these methods to other time series with different water consumption profiles, to see if the results remain similar. Here we only used the information of the water flow and the time when this observation was collected. It could help to achieve better results if we have access to some variables that could explain the profile of the consumption. For example climate variables, the number of consumers at each time, among others that could affect directly the water consumed. Neural networks could also be used to analyze these time series, the results could also be compared to see if they provide better predictions. All the time series used in this paper were static, it could be useful to create models capable of dealing with dynamic series and provide predictions in real time. Regarding the methods used, it is important to rewrite the built function in R in order to optimize the parameters by itself, this would for sure reduce the computation time needed to get the model and would conduct to better results. It is important to add prediction intervals to theirs computation, it would help the company to analyze the results and define a water demanding strategy for the predicted timestamps. In addiction, regarding the irregular time series it would be interesting to create a function in R that produced a decomposition of the series like the one that already exists for equally spaced time series.

References

[1] M. Aldrin and E. Damsleth. Forecasting non-seasonal time series with missing observations. *Journal of Forecasting*, 8:97–116, 1989.

[2] T. Cipra. Robust exponential smoothing. *Journal of Forecasting*, 11(1):57–69, 1992.

[3] T. Cipra. Exponential smoothing for irregular data. *Applications of Mathematics*, 51:597–604, 2006.

[4] A. Eckner. A framework for the analysis of unevenly spaced time series data. Working paper, July 2014.

[5] A. Eckner. A note on trend and seasonality estimation for unevenly spaced time series. Working Paper, June 2018.

[6] A. Eckner. Algorithms for unevenly spaced time series: Moving averages and other rolling operators. Working Paper, April 2019.

[7] T. Hanzák. Methods for periodic and irregular time series. Master's thesis, Charles University, 2014.

[8] T. Hanzák and T. Cipra. Exponential smoothing for irregular time series. *Kybernetika*, 44:385–399, 2008.

[9] R. H. Jones and P. V. Tryon. Continuous time series models for unequally spaced data applied to modeling atomic clocks. *SIAM Journal on Scientific and Statistical Computing*, 4(1):71–81, January 1987.

[10] T. Ratinger. Seasonal time series with missing observations. *Applications of Mathematics*, 41:41–55, 1996.

[11] J. T. T. Cipra and A. Rubio. Holt-winters method with missing observations. *Management Science*, 41:174–178, 1995.

[12] D. J. Wright. Forecasting data published at irregular time intervals using extension of holt's method. *Management Science*, 32:499–510, 1986.