Motifs In Time Series For Prediction

A naïve approach compared to ARIMA

Nertila Ismailaja
Programme Analyst
Armundia Factory
Tirana, Albania
n.ismailaja@armundiafactory.com

***Abstract*—Time series is a subject that includes two key factors - observations and time. It is obvious that observations are time-dependent. Another interest field is motif discovery, composed solely by time series subsequences. During this time, loads of similarity measures have been presented. In their article, Dhamo et al [5] concluded that the best performance according to the quality in motif discovery was achieved by Chouakria's index with CID (Chouakria's index, proposed by Chouakria et al [4] and CID is proposed by Batista et al [2]). The following step is to use this distance and other time series features to make predictions. Various tests are made over time series with high level of complexity. The results achieved by this approach are compared to ARIMA models. All the tests are made in R.**

Keywords—motif discovery;ARIMA;time series; forecasting; Chouakria with CID; R

I. Introduction

Motif discovery has been widely exploited in Biological sciences, to detect anomalies in heart beatings, pressure, etc. The effort was to deal with a large amount of data (known as big data). The first approach was made by reducing data dimensions' [1][3]. Some methods were Discrete Fourier Transformation[1], Discrete Wavelet Transformation[3], Piecewise Linear Approximation [12], etc. The first scientists to formally use motif discovery in time series, were Agrawal et al [3], Lin et al [6], Mueen[7][8][9][10], etc. Since then, several tactics have been proposed. The studies have always followed the similarity search throughout the time series, regarding to the problem. The most recent method is *ε-queries*, proposed as a probability approach to motif discovery. An expansion of motif discovery is to compare time series measures to each-other, based in four criterions - algorithm complexity, number of discovered motifs, accuracy and quality. In their article, Dhamo et al. [4] concluded that Chouakria index with CID gave better results than CID alone. Taking in consideration that Chouakria index with CID is a well-performing similarity measure, we move on in the following step- applying motif discovery in forecasting. There are used several well-known time series, for each of which is created a model, and then compared to the model given by ARIMA (to construct the model was used a predefined package of R, "forecast"). According to the results, by using as norm error in forecasting, we can deduce that our model provides better performance.

II. Time Series

A. Basic Concepts

A time series A time series may be defined as a collection of data, where time is a relevant component.

Definition 1 A time series T is an ordered collection of data, observed in n - regular intervals of time $[T1, T2, \dots, Tn]$.

Definition 2 A motif M of length m in a time series T of length n is a subsequence of T which repeats itself in T.

Definition 3 In *ε-query* search for similarity between two time series' subsequences $P = [P_1, P_2, \dots, P_m]$ and $Q = [Q_1, Q_2, \dots, Q_m]$ of length m, in a time series T of length n, P and Q are considered similar if the distance between P and Q is laid in an interval with absolute error equal to ε .An illustration of a motif is given in Fig. 1., with dataset unemp.

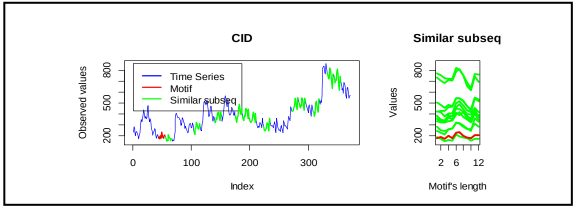


Fig.1. Algorithm for motif discovery

B. Distances Used For Time Series

While comparing two notions, it is indispensable to use a criterion for comparison, in order to claim whether these notions mean or not the same. In time series, this criterion is undoubtedly the similarity measure, which is based in the concept of the distance.

Definition 4 A distance is a function that complies with the following properties:

- Identity: $\forall x, y \in R, d(x,y) = 0 \leftrightarrow x = y$
- Symmetry: $\forall x, y \in R, d(x,y) = d(y,x)$
- Non-negativity: $\forall x, y \in R, d(x,y) \geq 0$
- Transitive: $\forall x, y \in R, \exists z \in R | d(x,y) \leq d(x,z) + d(z,y)$

Definition 5 A similarity measure is a function that disobeys at least one of the conditions to be distance.

In other words, a similarity measure is able to measure the similarity between two time series' subsequences. Reported as a key factor in the quality and quantity in motif discovery, there are several similarity measures. In contrary to its' wide usage, Euclid distance provides dissatisfactory results if it is used as a similarity measure.

Definition 6 Euclid distance between two subsequences with length *m*, $P = [P_1, P_2, \dots, P_m]$ and $Q = [Q_1, Q_2, \dots, Q_m]$, is the index, measured as below:

$$d_{Euc}(P,Q) = \sum_{i=1}^{m}(P_i - Q_i)^2 \tag{1}$$

As we mentioned above, the most performing similarity measure is Chouakria's index with CID. In this case, it is therefore necessary to explain what is CID, and then to give the definition of Chouakria's index.

Definition 7 CID distance between two sub-sequences with length *m*, $P = [P_1, P_2, \dots, P_m]$ and $Q = [Q_1, Q_2, \dots, Q_m]$, is the index, measured as below:

$$d_{CID}(P,Q) = d_{Eucl}(P,Q)\frac{\max\{CE(P), CE(Q)\}}{\min\{CE(P), CE(Q)\}} \tag{2}$$

, where:

$$CE(P) = \sum_{i=1}^{m-1}(P_i - P_{i+1})^2 \tag{3}$$

This similarity measure is used for nonlinear time series subsequences. One main feature of CID is that, in case of two nonlinear subsequences P, Q of length m:

$$d_{CID}(P,Q) > d_{Eucl}(P,Q) \tag{4}$$

Moreover, the closer to 1 is the fraction $\frac{\max\{CE(P), CE(Q)\}}{\min\{CE(P), CE(Q)\}}$, the more similar is the behavior of the two subsequences.

Another similarity measures is Chouakria's index. This measure is composed by two factors, one of which is responsible for behavior and the other for proximity of values[]. Chouakria's index is measured, as below:

Definition 8 Chouakria's index between two sub-sequences with length *m*, $P = [P_1, P_2, \dots, P_m]$ and $Q = [Q_1, Q_2, \dots, Q_m]$, is the index, measured as below:

$$d_{CID}(P,Q) = \frac{2}{1+e^{k*\delta(P,Q)}} * COR_t(P,Q) \tag{5}$$

, where:

$$COR_t(P,Q) = \frac{\sum_{i=1}^{m-1}(P_i - P_{i+1})(Q_i - Q_{i+1})}{\sqrt{\sum_{i=1}^{m-1}(P_i - P_{i+1})^2}\sqrt{\sum_{i=1}^{m-1}(Q_i - Q_{i+1})^2}} \tag{6}$$

and $k \in R^+$; and $\delta(P,Q)$ may be Euclid, etc. In their article, Dhamo et al. [] proposed CID as distance $\delta(P,Q)$ and $k = 2$.

III. MODELLING TIME SERIES

Time series is a novel concept, which enrolls two main concepts - determinism and randomness. A time series is composed by many parts, such as trend, seasonality, cycles, etc. After the detection of these components in a time series, the remains are considered a random part. In other words, a time series *T* of length *n*, can be decomposed, as below:

$$T_n = t_n + c_n + s_n + \varepsilon_n \tag{7}$$

, where t_n is trend, c_n is cycle, s_n is seasonal variation and ε_n is noise. Usually a time series can be decomposed according to these features. The only characteristic to be modeled is the noise, which is a random variable.

A. ARIMA Models

As cited above, a time series is a regular collection of data. An ARIMA model is composed by implementing other forecasting models, such as AR, MA and the operator Δ. Let us first define the key concepts of a stochastic process.

Definition 9 Let $\{\varepsilon_t, t \in T\}$ be a stochastic process. $\varepsilon_t \sim WN(0, \sigma^2)$ is considered to be a white noise, if it obeys the following rules:

$$\begin{cases} E(\varepsilon_t) = 0 \\ E(\varepsilon_t \varepsilon_s) = \begin{cases} \sigma^2, s = t \\ 0, s \neq t \end{cases} \end{cases} \tag{8}$$

Definition 10 Let $\{Y_t, t \in T\}$ be a stochastic process and $\varepsilon_t \sim WN(0, \sigma^2)$. An AR(p) model is constructed, as below:

$$\varepsilon_t = \sum_{i=0}^{p} \varphi_i Y_{t-i} \tag{9}$$

, where $\varphi_i, i = \overline{1,p}$ are coefficients defined dynamically.

Definition 10 Let $\{Y_t, t \in T\}$ be a stochastic process. Let $\omega_i \sim WN(0, \sigma^2), i = 0,1,2,\dots$. A MA(q) model is constructed, as below:

$$Y_t = \sum_{i=0}^{q} \psi_i \omega_{t-i} \tag{10}$$

, where $\psi_i, i = \overline{1,p}$ are coefficients defined dynamically.

Logically, a ARMA(p,q) model, is constructed, as below:

Definition 12 A stochastic process $\{Y_t, t \in T\}$ is said to be ARMA(p,q) model (autoregressive moving average of order (p,q)) if it can be expressed as:

$$Y_t = \mu_t + \sum_{i=0}^{p} \varphi_i Y_{t-i} + \sum_{i=0}^{q} \psi_i \omega_{t-i} \tag{11}$$

, where $\mu_t, \varphi_i, i = \overline{0,p}$ and $\psi_i, i = \overline{0,p}$ are constants and ω_t is white noise.

An important parameter in ARIMA(p,d,q) is also the difference operator (Δ), defined as below:

Definition 13 Let $\{Y_t, t \in T\}$ be a stochastic process. The Δ^d operator is though defined, as:

$$\Delta^d = Y_t - Y_{t-d} \tag{12}$$

, where $d = 1,2, \dots$

Having all the necessary concepts, we can introduce ARIMA concept.

Definition 14 The stochastic process $\{Y_t, t \in T\}$ is considered to be an ARIMA(p,d,q) process only if the process

$$X_t = \Delta^d Y_t \qquad (13)$$

is an ARMA(p,q) process.

Of course that there are several ways to determine the most efficient values of p,q and d. But this will be discussed in the following sections.

B. The Proposed Approach

When trying to detect a motif in a time series, it is agreed that there is a pattern that repeats itself in time series, independently from observation. An example is given in Fig. 2.

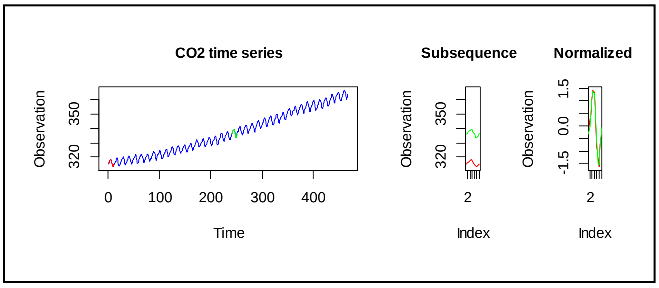


Fig.2. The relation between time series' sub-sequences (time series co2)

It is obvious the relationship that exists between the subsequences of length 12. The relation is made visible by the normalization of these subsequences, where is impossible to distinguish subsequences from each-other.

Definition 15 Normalization of the time series T with length n, is the time series T', defined as:

$$T' = \frac{T - mean(T)}{sd(T)} \qquad (14)$$

If there is a such strong relation between two subsequences $P = [T_i, T_{i+1}, \dots, T_{i+m-1}]$ and $Q = [T_j, T_{j+1}, \dots, T_{j+m-1}]$ with length m, from the time series T, it is presumable that the relation would be of the same strength even if the length of the subsequence is greater. The situation is shown in Fig. 3.

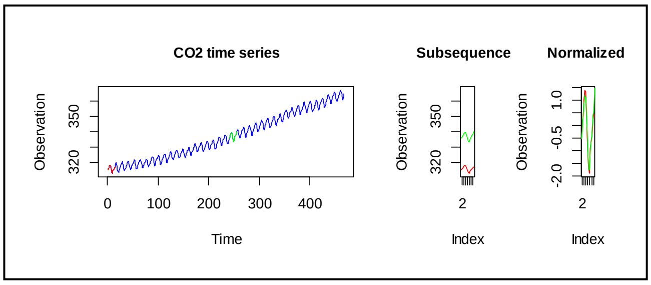


Fig. 3. Widening the motif's length

It is clear from Fig. 3. that, even the length of the subsequences has been enlarged, from 12 to 15, the difference is still small.

C. Reasoning to Reach to Forecasting

When we say that there is a motif in a time series, we assume that there is a large similarity between different subsequences of the same time series. Moreover, these subsequences are generally spread throughout all the time series. Given a fixed length m as motif's length, we are able to find, according to Brute-Force algorithm, all patterns in the time series.

```
Brute -force algorithm
Brute_force=function(T,m,epsilon)
    1.#n=length(T)
    2.threshold=min{d(Ai,Bj)}, Ai=T[i : (i+m-1)],i=1: (n-m+2) and
    j=(i+1): (n-m+2)
    3.motif_index=argmax{d(Ai,Bj)<threshold+epsilon}
    4.motif=T[motif_index: (motif_index+m-1)]
    5. similar_seq=T[where(d(motif,Bj)<threshold+epsilon)], Bj=T[j :
    (j+m-1)] and j>motif_index
end function
```

In contrary to this algorithm, in order to predict, there are some changes being made. Those changes are reflected not only in code, but even in structure:

```
Short-term algorithm
Short_term=function(T,m)
    1.Keep the last subsequence of length m constant (also considered as
    motif, $T[(n − m + 1) : n]$.).
    2.Choose one suitable distance for time series (Chouakria's index)
    3. Normalize time series' subsequences, by using (8).
    4.Detect the strongest relationship between our motif and other
    subsequences of our time series. Keep track of the initial index
    where the strongest relationship is proven.
    5.Create a dependency factor that is being used during the
    prediction.
end function
```

Suppose that we require to use a subsequence of length m in our time series T of length n. Moreover, is found out that the most approximate subsequence is stated in position j. In this case, the dependency factor would be calculated, as below:

$$dep_{fact} = \frac{T[j+m]}{T[j+m-1]} \qquad (15)$$

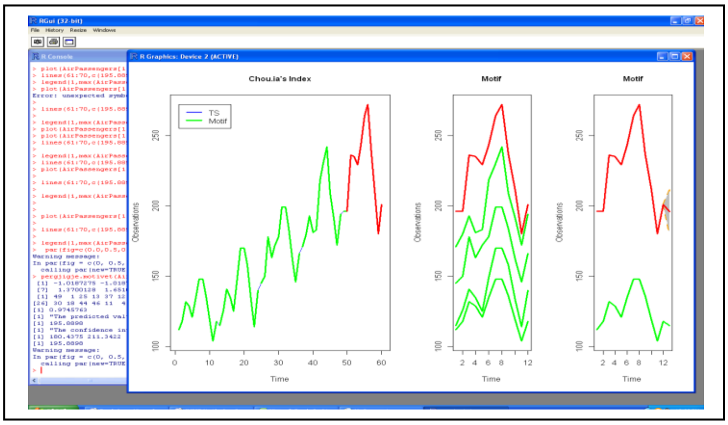
Our prediction is $T[n+1]$, which is calculated as:

$$T[n+1] = dep_{fact} * T[n] \qquad (16)$$

We define $a = T[(n-m+1):(n+1)]$. we can create a 95% confidence interval for our prediction, as below:

$$CI(T[n+1]) =]E(a) - 1.96\frac{sd(a)}{\sqrt{m+1}}, E(a) + 1.96\frac{sd(a)}{\sqrt{m+1}} \qquad (17)$$

An example of our algorithm in R, in time series AirPassengers, is given in Fig. 4.



```
> plot(AirPassengers[1
> lines(61:70,c(195.88
> legend(1,max(AirPass
> plot(AirPassengers[1
Error: unexpected symb
>
> lines(61:70,c(195.88
>
> legend(1,max(AirPass
> plot(AirPassengers[1
> plot(AirPassengers[1
> lines(61:70,c(195.88
>
> legend(1,max(AirPass
> lines(61:70,c(195.88
> plot(AirPassengers[1
>
> lines(61:70,c(195.88
>
> legend(1,max(AirPass
>
>
> plot(AirPassengers[1
>
> lines(61:70,c(195.88
>
> legend(1,max(AirPass
>  par(fig=c(0.0,0.5,0
Warning message:
In par(fig = c(0, 0.5,
  calling par(new=TRUE
> pergjigje.motivet(Ai
 [1] -1.0187275 -1.018
 [7]  1.3700128  1.651
 [1] 49  1 25 13 37 12
[26] 30 18 44 46 11  4
[1] 0.9745763
[1] "The predicted val
[1] 195.8898
[1] "The confidence in
[1] 180.4375 211.3422
[1] 195.8898
Warning message:
In par(fig = c(0, 0.5,
  calling par(new=TRUE
> 
```

Fig.4. Snapshot of the proposed approach

D. Forecasting in a Long-term Period

In the previous paragraph, we constructed an algorithm to predict the next observation, based on previous ones. What is more, there is a possibility to find a confidence interval for the provided prediction. But, in practice, there is a long-term necessity for prediction. In order to get to a long-term forecasting, we construct the following algorithm, as below:

```
Long-Term Algorithm
Long_Term=function(T,m,k)
    #n=length(T); k is the number of lags we want to forecast
    for (i in seq(1,k)){
    prediction=Short_Term(T,m)
    T=c(T,prediction)
    }
end function
```

According to this algorithm, we can keep track of the original time series and the forecasts provided. An example is provided below, in Fig. 5.

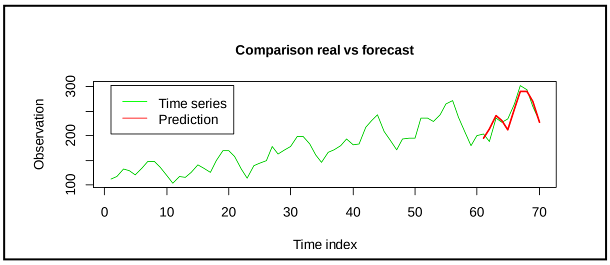


Fig.5. Forecasting in long-term period

In Fig.5. is obviously seen the proximity in values and even in shape. This means that the model that is built is strong.

IV. ARIMA VS PROPOSED METHOD

In many forecasting models, such as AR, MA, ARIMA, etc, the basis of fitting the best parameters is *low error in curve fitting*. The lower the error it is, the better are the parameters.

Definition 16 Given T a time series of length n, and T' the ARIMA fitted model. The error in curve fitting is called the vector, as below:

$$\varepsilon = T - T' \tag{18}$$

In practice, most commonly used is the Sum Square Error ($\varepsilon'\varepsilon$). There are also other measures that supply with information about the quality of the constructed model, such as AIC (Akaike Information Criterion) or BIC (Bayes Information Criterion).

Definition 11 Given T a time series of length n, and T' the ARIMA fitted model. AIC is measured, as below:

$$AIC = \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + \frac{2(k+1)}{n} \tag{19}$$

, where $e_i^2 = (T_i - T_i{}')^2$ and k is the degree of freedom of T.

Definition 12 Given T a time series of length n, and T' the ARIMA fitted model. BIC is measured, as below:

$$BIC = \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + \frac{(k+1)*\ln(n)}{n} \tag{20}$$

, where $e_i^2 = (T_i - T_i{}')^2$ and k is the degree of freedom of T.

Generally, between AIC and BIC criterions, the most useful is BIC, which is considered to be more accurate.

A. Detecting Error in Forecast in the Proposed Model

In the proposed algorithm, we mentioned that the method does not create a single model - it does not create parameters. Our model is a dynamic model, which, in a long-term period, changes constantly. This means that AIC or BIC criterion is inapplicable.

In order to be able to proof which model is better, and why, we use the concept of *error in forecasting*. This means that we keep a certain percentage of values unused (the last observations), in order to detect which algorithm provides a smaller Sum Square Error in Predictions. The number of predictions is kept constant, 10, in any case that we studied. This is done in order to prevent reducing the amount of information in disposal.

In Table 1, there are some examples of errors made during the forecast, for several time series.

TABLE I. COMPARISON OF TWO METHODS

Time series	**Error in forecasting**	
	ARIMA	***Prop. approach***
star	**4184.963**	4245.082
prodn	**16.14206**	18.56693
rosewine	Inf	**768.4422**
unemp	15006.22	**4095.185**
penguin	596901.4	**321.071**
Al_births	Inf	**3489920**
hsales	671.1774	**274.1702**
AirPassengers	281852.9	**167063.2**

ARIMA vs Proposed approach

In Table 1, is obvious the advantage of our approach in relation to ARIMA. In 2 cases, ARIMA could not give an appropriate model – AIC criterion could not be approximated. In our trials, resulted that in 66.7% of the cases, our approach provided better performance than ARIMA's model. In this percentage, is also included the 12.5% when ARIMA failed to provide a suitable model.

An example of how much differ the real results from the ones provided by ARIMA, is given in Fig.6.

The time series taken in consideration is *penguin,* while the length of the motif is 7.

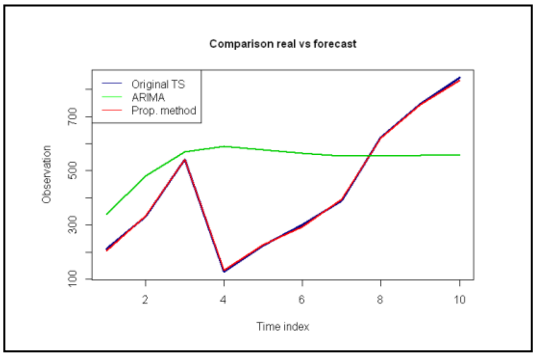


Fig.6. ARIMA, the prop. method and real values

It is clearly visible that the shape and the values gained by the proposed method are more compatible with the real ones. Whereas ARIMA's curve of prediction is far more unrealistic. In many cases, ARIMA provided central values, very close to a linear curve. Whereas the proposed method provides different shapes, a nonlinear vector of forecasts.

B. *Detecting factors that influence results*

In order to define where this result comes from, so, whether it is because of the complexity of the time series, we use a complexity measure for time series' complexity-fluctuations[9].

Definition 17 Given T a time series of length n, fluctuations are measured, as below:

$$fluct(T) = \frac{1}{n-1}\sum_{i=1}^{n-1}(T_i - T_{i+1})^2 \qquad (21)$$

In Fig.7. is shown the fluctuation of some time series.

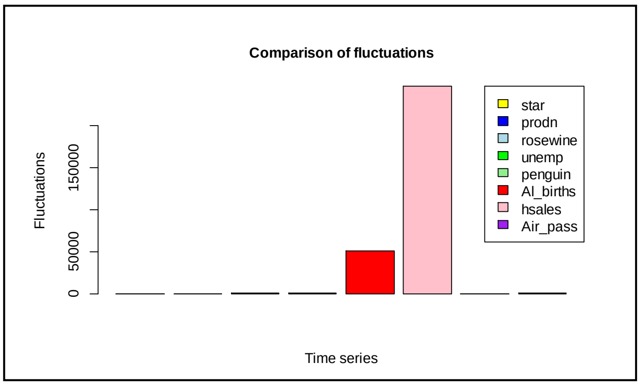


Fig.7. Fluctuation of some time series

According to the traditional definition to the correlation between two factors, as below:

Definition 18 The correlation between two random variables X and Y with length n is the index, measured as below:

$$COR(P,Q) = \frac{\sum_{i=1}^{n}(P_i-\bar{P})*(Q_i-\bar{Q})}{\sqrt{\sum_{i=1}^{n}(P_i-\bar{P})^2}\sqrt{\sum_{i=1}^{n}(Q_i-\bar{Q})^2}} \qquad (22)$$

it results that, in both cases, the relation between fluctuation and error in forecasting is strong (in each case, greater than 99%). This means that there is a strong dependency between these two factors.

Another important factor that influences the forecast, might be the quantity of the data. It is naively deduced that the longer the time series is, the smaller will be the error. There have been made various tests, by increasing the amount of data. It has been chosen a high-complexity time series, such as *hsales*. In Fig.8. we see what happens with the error in forecasting when this time series is enlarged, with fixed length of the time series m=12 and m=9.

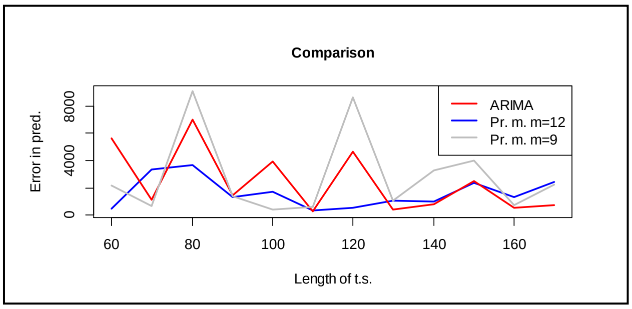


Fig.8. Error in forecasting for different lengths

In 50% of the cases, ARIMA (the proposed method with m=12) were better than the other. In a low percentage, the proposed method with m=9 is more performant than the others. What is more, we notice that there is no defined trend of the error in forecasting. This means that, if we increase the amount of data, we cannot presume whether the error will decrease or not.

An important parameter during our method would be the length of the motif, m. The time series selected for this evaluation is again *hsales*, due to its' complexity. We keep a fixed length of the time series- n=170. An example of this algorithm is shown in Fig.9.

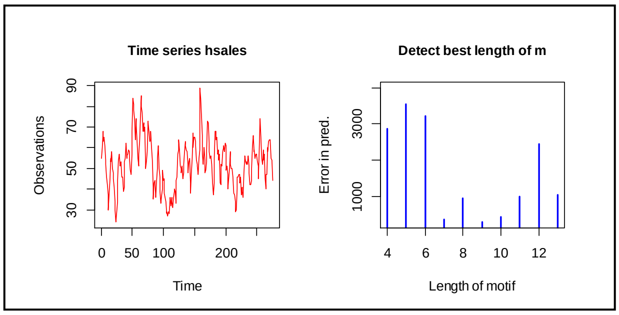


Fig.9. a) hsales b) Detecting the best value of m

We can see that the lower error in prediction is made for m=9, equal to the periodicity of the time series. But this contradicts the upper results, where the proposed approach with m=12 was better than the proposed approach with m=9. This means that, in order to get better results, the length of the motif should be defined according to a careful study of periodicity, variations, etc.

Acknowledgment

I would like to thank Ms. Ilda Drini, employee at Department of Financial Stability and Statistics, for all the help during the creation of this method.

References

All time series that are used in this article, can be found in the next websites:

http://new.censusatschool.org.nz/resource/time-series-data-sets-2013/

www.instat.gov.al
https://datamarket.com/data/list/?q=provider%3Atsdl
http://www.cs.ucr.edu/~eamonn/discords/
http://vincentarelbundock.github.io/Rdatasets/datasets.html
http://cran.r-project.org/web/packages/astsa/index.htm

[1] Agrawal, R., Faloutsos, C., Swami, A., (1993): “Efficient similarity search in sequence databases,” in: Fourth International Conference on Foundations of Data Organization, D. Lomet, Ed., Heidelberg: SpringerVerlag, pp. 69–84

[2] Batista, G., Keogh, E. J. Tataw, O., de Souza, V., “CID: an efficient complexity-invariant distance for time series”

[3] Chan, K. & Fu, W. (1999). Efficient time series matching by wavelets. Proceedings of the 15 th IEEE International Conference on Data Engineering.

[4] Chouakria, A., D., Diallo A., Giroud F., (2007) “Adaptive clustering of time series”. International Association for Statistical Computing (IASC), Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal

[5] Dhamo, E., Ismailaja, N., Kalluçi, E., (2015): “Comparing the efficiency of CID distance and CORT coefficient for finding similar subsequences in time series”, Sixth International Conference ISTI, 5-6 June.

[6] Lin, J., Keogh, E. , Lonardi, S. and Patel, P. (2002): “Finding Motifs in Time Series”, in Proc. of 2nd Workshop on Temporal Data Mining

[7] Mueen A., Keogh, E., J., (2010A): “Online discovery and maintenance of time series motifs”. KDD, pg. 1089-1098

[8] Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B., (2009A) “Exact Discovery of Time Series Motifs”, SDM, pg. 473-484

[9] Mueen A., Keogh, E., J., Bigdely- Shamlo N., (2009): “Finding Time Series Motifs in Disk-Resident Data”. ICDM, pg 367-376

[10] Mueen A., (2013):” Enumeration of Time Series Motifs of All Lengths”. ICDM,pg. 547-556

[11] Yan, Ch., Fang, J., Wu, L., Ma, Sh., (2013): “An Approach of Time Series Piecewise Linear Representation Based on Local Maximum Minimum and Extremum”, Journal of Information & Computational Science 10:9 2747–2756

[12] Yi, B.-K., Faloutsos Ch., (2000): “Fast time sequence indexing for arbitrary Lp norms”. In The VLDB Journal, pages 385–394.