Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 17 Issue 1 Version 1.0 Year 2017
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Loss of Fitting and Distance Prediction for Fixed vs Updated ARIMA Models

By Livio Fenga
University of California San Diego

Abstract- In many cases, it might be advisable to keep an operational time series model fixed for a given span of time, instead of updating it as a new datum becomes available. One common case, is represented by model–based deseasonalization procedures, whose time series models are updated on a regular basis by National Statistical Offices. In fact, in order to minimize the extent of the revisions and grant a greater stability of the already released figures, the interval in between two updating processes is kept "reasonably" long (e.g. one year). Other cases can be found in many contexts, e.g. in engineering for structural reliability analysis or in all those cases where model re–estimation is not a practical or even a viable options, e.g. due to time constraints or computational issues. Clearly, the inevitable trade–off between a fixed models and its updated counterpart, e.g. in terms of fitting performances, out–of–sample prediction capabilities or dynamics explanation should be always accounted for.

Keywords: *ARIMA models, model stability, model fitting, time series distances measure, time series prediction.*

GJSFR-F Classification: *MSC 2010: 97K80*

Strictly as per the compliance and regulations of :

Loss of Fitting and Distance Prediction for Fixed vs Updated ARIMA Models

Livio Fenga

Abstract- In many cases, it might be advisable to keep an operational time series model fixed for a given span of time, instead of updating it as a new datum becomes available. One common case, is represented by model–based deseasonalization procedures, whose time series models are updated on a regular basis by National Statistical Offices. In fact, in order to minimize the extent of the revisions and grant a greater stability of the already released figures, the interval in between two updating processes is kept "reasonably" long (e.g. one year). Other cases can be found in many contexts, e.g. in engineering for structural reliability analysis or in all those cases where model re–estimation is not a practical or even a viable options, e.g. due to time constraints or computational issues. Clearly, the inevitable trade–off between a fixed models and its updated counterpart, e.g. in terms of fitting performances, out–of–sample prediction capabilities or dynamics explanation should be always accounted for. This paper is devoted at presenting a procedure for the prediction of the loss in terms of fitting ability of a fixed model of the type autoregressive integrated moving average versus its updated version – according to a suitable quadratic cost function – and at giving a quantitative measure of the discrepancy between them. Being the updating frequency customizable, the presented approach can also be employed for simulations purposes, according to the updating intervals, the degree of complexity of the chosen model and the available computing resources. Finally, an empirical experiment involving both computer simulated and macroeconomic time series will be presented and the related outcomes discussed.

Keywords: ARIMA models, model stability, model fitting, time series distances measure, time series prediction.

I. Introduction

There are many reasons which might justify the choice of leaving a time series model – once correctly estimated and tested – unchanged for a certain time span, even when its performances, as expected, tend to deteriorate. The extent to which such a degradation can be considered acceptable, heavily depends on the specificity and the target a given model is built for. Under pre-specified regularity conditions, e.g. in terms of stability in the model's outcomes or of the underlying Data Generating Process (DGP), the benefits of using the same model for a given period of time are mainly related to two important factors: the need of a greater stability of the model outputs and to keep the computational time within "reasonable" limits and within the limits of the available computing resources. As for the first point, its relevance is evident in the case of statistical providers (e.g. national and supernational statistical offices), which constantly check past data for consistency with the most recent official releases. It is not uncommon, in fact, that in the attempt of capturing new features exhibited by the time series at hand (which might have had an irrelevant impact on past data or even

Author: Istat and University of California San Diego. e-mail: fenga@istat.it

gone undetected) the model is subjected to too frequent updating procedures involving, for example, the structure of the vector of parameters, the introduction of auxiliary variables (e.g. of the type dummy) or even the inference procedures. However, such interventions can jeopardize the coherence with data already released, validated and, therefore, employed in many type of official and unofficial analysis. This is, for example, the case of the model–based signal extraction techniques, which can be carried out by two widely employed deseasonalization methods, i.e. X-12-ARIMA [1] and TRAMO-SEATS [2]. As it is well known, they might generate, as a inevitable "byproduct", the undesirable phenomenon of the revisions, which is due to the inclusion in the data set of each and every new observation as it becomes available. In more details, at the current end of a time series, it is not possible to use symmetric filters to estimate the trend because of the end point problem. Instead, asymmetric filters are used to produce provisional trend estimates. However, as more data becomes available, it is possible to recalculate the trend using symmetric filters and improve the initial estimates. As expected, the impact of the revisions is more noticeable both in the period immediately preceding the inclusion of the new data and the corresponding period one seasonal lag prior. This problem has attracted a great number of researchers, triggering a still ongoing discussion on the different methods and procedure to deal with it. In particular, the problem has been discussed by[3], [4], [5] and, more recently, in [6]. Many other situations can require the use of a fixed model, e.g. when model's outputs must be provided under strict time limits – leaving not enough room for building and test a new model – or the nature of the Data Generating Process (DGP) under investigation suggests the changes in the model only reflect temporary phenomena, for instances related to outlier of the type temporary change, influential data, survey issues (e.g. unexpected amount of missing data). Another common scenario pertains the assessment of model lack of fitting, in order to monitor the stability of the underlying DGP. In this perspective, valuable insights can be gained in economics, e.g. to detect the changes occurring over time – as well as their starting points – in the case of key variable, such as the industrial production or the inflation indexes. Other important applications are related to on–line monitoring activities, e.g. for safety level assessment of structures – such as bridges, dams, TV towers – under standard as well as abnormal conditions, e.g. of the type of those induced by automotive traffic, temperature changes, wind, distant earthquakes, landslides (for a review of the most used methods the reader is referref to [7] [8]). For example, in [9] the modeling of the vibration signals originating from a bridge has been performed using a model of the class ARIMA, whereas mode-based damage identification techniques have been discussed in [10]. This framework identifies a class of problems of the type "inverse", as their design envisions a "baseline" model, whose structure identification and parameter inference procedures, however, usually inject a not negligible amount of uncertainty in the system under investigation [11]. In order to control for such a source of uncertainty, the input series has been modelled here assuming a DGP of the type autoregressive integrated moving average (ARIMA) [12], which in general can guarantee a good level of robustness and, unlike other methods, does not assume any particular pattern in the historical data. In addition, other being a plausible hypothesis satisfactorily adopted for many real-life phenomena (e.g. in economics, physics or engineering) this class of models enjoys a well established theoretical framework and that many routines are nowadays available free of charge for its efficient estimation. The proposed procedure uses an ad hoc distance function in conjunction with a suitable quadratic loss function and an extrapolation method. In

Ref

1. David F Findley, Brian Monsell, Mark Otto, William Bell, and Marian Pugh. Towards x12 arima. Technical report, mimeo, Bureau of the Census, 1992.

particular, in the Empirical Section two different distance metrics – i.e. the Complexity Invariant and the Normalized Integrated Periodogram distances – and two extrapolation methods – i.e. of the type polynomial regression and double exponential smoothing – will be considered.Clearly, the ARIMA assumption can be easily relaxed and a different type of model used, without changing the structure of the proposed framework, provided that a suitable metric for the estimation of the distance between models is correctly chosen. Consistently, in the empirical section, two model–free distances are applied. The proposed procedure might be also a useful tool for balancing model fitting, prediction performances and stability of the outcomes.

II. The Method

Throughout the paper, the time series of interest is intended to be a real–valued, uniformly sampled, sequence of data points of length T, denoted as

$$x_t := \{(x_t)_{t\in\mathbb{Z}^+}^{T}\}, \tag{1}$$

whereas its predicted values at horizon h are formalized as follows:

$$x_t(h) = \{(x_t)_{t\in\mathbb{Z}^+}^{T+h}; \quad h = 1,2,\dots,H\}. \tag{2}$$

An arbitrary, length $\mathcal{H} \in \mathbb{Z}^+$, windows is chosen as the time span in which a given model structure $M^{\bullet}$ estimated conditional to the full information available at the time $t-1$, i.e. $M^{\bullet}(t+h) = |\mathcal{I}_{t-1}$, is kept fixed for $\mathcal{H}$ times until an upper bound $\bar{\mathcal{H}}$ is reached, i.e. $t+1, t+2, \dots, t+\bar{\mathcal{H}}$. This model is formalized as follows: $M^{\circ}_{\mathcal{H}}(t+h) = |\mathcal{I}_{t-1-\mathcal{H}} \quad \mathcal{H} = 1,2,\dots\bar{\mathcal{H}}$.

Consistently, the predicted values obtained by $M^{\circ}(t+h)$ and $M^{\bullet}(t+h)$ are respectively denoted by $y^{\circ}(t+h)$ and $y^{\bullet}(t+h)$ therefore assuming i.e. $y_t^{\bullet}\sim ARIMA(p_0,d_0,q_0); \quad \mathcal{H}=1$ and $y_t^{\circ}\sim ARIMA(p_0,d_0,q_0); \quad \mathcal{H}=2,3\dots,\bar{\mathcal{H}}$, we will have that $y_t^{\bullet}(h) \equiv y_t^{\circ}(h) \Leftarrow \mathcal{H}=1$ for each horizon considered $h = 1,2,\dots,H$.

a) The underlying stochastic process and the distance measure adopted

The proposed procedure assumes the input time series (1) to be a realization of a DGP of the class ARIMA. Let x_t be a realization of a real 2^{nd} order stationary DGP, with mean μ. It is said [12] to admit a Autoregressive Moving Average representation of order p and q – i.e. $x_t\sim ARMA(p,q)$, with $(p,q)\in\mathbb{Z}^+$ – if for some constant $\phi_1\dots\phi_p$, $\theta_1\dots\theta_q$, it is:

$$\sum_{j=0}^{p}\phi_j(X_{t-j}-\mu) = \sum_{j=0}^{q}\theta_j\alpha_{t-j} \tag{3}$$

Eqn. (3) is valid under the following assumptions: *a)* $\phi_0=\theta_0=1$; *b)* $E\{\alpha(t)|\mathcal{S}_{t-1}\}=0$; *c)* $E\{\alpha^2(t)|\mathcal{S}_{t-1}\}=\sigma^2$; *d)* $E\alpha^4(t)<\infty$; *e)* $\sum_{j=0}^{p}\phi_j z^j \neq 0$, $\sum_{j=0}^{p}\theta_j z^j \neq 0, |z|\leq 1$, where $\mathcal{S}_t$ denotes the sigma algebra induced by $\{\alpha(j), j\leq t\}$ and $\sum_{j=0}^{p}\phi_j z^j$ and $\sum_{j=0}^{p}\theta_j Z^j$ are assumed not to have common zeros. When needed, x_t can be transformed into a stationary process by differencing it $d\in\mathbb{Z}^+$ times. The order of integration, denoted as $I(d)$, enters formally in the ARIMA scheme, i.e. $x_t\sim ARIMA(p,d,q)$, so that using the back-shift operator L, i.e. $LX_t = X_{t-1}$ (therefore

12. George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

$L^n X = X_{t-n}$) and the difference operator $\nabla^d X_t = (1-L)^d X_t d = 0,1,\ldots D$, the ARIMA model can more synthetically be expressed as

$$\nabla^d (x_t - \mu) = \frac{\theta(L)}{\phi(L)} \alpha_t, \tag{4}$$

with $\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p$; $\theta_q(L) = 1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_q L^q$, and difference operator applied d times until stationarity is reached. Here ϕ, θ and α_t are, respectively, the autoregressive and moving average parameters. The term α_t is the white noise sequence with mean $\mu = 0$ and variance $\sigma^2 < \infty$. The estimation of (4) is possible only if the stationary and invertibility conditions are satisfied for both the autoregressive and moving average polynomials respectively, that is when $\phi_p(L)\theta_q(L) = 0$ has roots lying outside the unit circle. On the other hand, here the estimation of the ARIMA order $(\hat{p}, \hat{d}, \hat{q})$, is based on the Akaike Information Criterion (AIC) [13], which is defined as $-2\text{maxlog}(L(\widehat{\boldsymbol{\theta}}|y)) + 2K$, with K the model dimension and $(L(\widehat{\boldsymbol{\theta}}|y))$ the log–likelihood function. The related selection strategy adopted, called $MAICE$ (short for Minimum AIC Expectation) [14], is a procedure aimed at extracting, among the set of the candidate models, the order $(\hat{p}, \hat{d}, \hat{q})$ satisfying:

$$(\hat{p}, \hat{d}, \hat{q}) = \arg \min_{p \le p_0, d \le d_0, q \le q_0} AIC(p, d, q). \tag{5}$$

$MAICE$ procedure requires the definition of an upper bound for all the AR and MA parameters as well as for the difference operators, as a maximum order a given process can reach. This choice, unfortunately, is a priori and arbitrary. As already pointed out, two distance measures are considered in the present paper: the complexity invariant (CI) and the one based on the normalized integrated periodogram ($NIPER$). They are both model free and measure the distance between two series, say $\mathbf{Y}_t$ and $\mathbf{X_t}$ (1), the former exploiting a corrected version of the Euclidean distance whereas the latter on the basis of a normalized nonparametric spectral estimators.

The CI metric has been recently proposed in [15] and subsequently discussed in [16], as a correction factor of a given distance measure driven by the complexity difference between two time series. In this paper, the Euclidean Distance ED(x,y), between two time series x and y is considered. It is made invariant through the correction factor γ so that the distance is expressed as follows:

$$\delta^{ci}(x, y) = ED(x, y) \times \gamma(x, y).$$

Here, γ is expressed by $\gamma(x, y) = \frac{\max\{\widehat{\mathbb{C}}(x), \widehat{\mathbb{C}}(y)\}}{\min\{\widehat{\mathbb{C}}(x), \widehat{\mathbb{C}}(y)\}}$, with $\widehat{\mathbb{C}}$ defining the series' complexity estimation, i.e.

$$\widehat{\mathbb{C}} = \sqrt{\sum_{t=1}^{T-1} (x_t - x_{t+1})^2}. \tag{6}$$

Following [5], it has to be emphasized how the one formalized in (6) is only one of the possible complexity measures – as many others can be successfully employed – but nevertheless it is particularly suitable for the problem at hand being model–free, $\mathcal{O}(T)$ time complexity and $\mathcal{O}(1)$ space. The other distance measure considered, is the Normalized Cumulated Periodogram Based Dissimilarity which is based on the

Ref

15. Gustavo EAPA Batista, Xiaoyue Wang, and Eamonn J Keogh. A complexity-invariant distance measure for time series. In *SDM*, volume 11, pages 699–710. SIAM, 2011.

cumulative periodogram of the series and has been proposed by [17]. Given the periodograms of $\mathbf{Y}$ and $\mathbf{X}$, respectively defined as $I_{X_t}(\mu_k) = \frac{1}{T}|\sum_1^T X_t e^{-i\mu t}|^2$ and $I_{Y_t}(\mu_k) = \frac{1}{T}|\sum_1^T Y_t e^{-i\mu t}|^2$, computed at frequencies $\mu_k = \frac{1}{T}2\pi k; \quad k = 1,2,\ldots,\frac{T-1}{2}$, the Normalized Cumulated Periodogram Based Dissimilarity takes the form

$$\delta^{per}(\mathbf{Y}_t\mathbf{X}_\mathbf{t}) = \int_{-\pi}^{\pi} |F_{X_t}(\mu) - F_{Y_t}(\mu)| d\mu, \tag{7}$$

being $F_{X_t}(\mu_j) = C_{X_T}^{-1}\sum_{i=1}^{j} I_{X_t}(\mu_i)$, $F_{Y_t}(\mu_j) = C_{Y_T}^{-1}\sum_{i=1}^{j} I_{Y_t}(\mu_i)$, with $C_{X_t} = \sum_i I_X(\mu_i)$ and $C_{Y_t} = \sum_i I_Y(\mu_i)$. Following ([9]) the normalized version of (7) has been adopted, as the two functions Fs in all the simulations conducted show a strong tendency to intersect. Finally, the adopted quadratic loss function is the $RMSFE$ (Root Mean Square Forecast Error), computed on the test set Ts. Based on the L_2 −norm, this metric is massively employed in the performance assessment stage of time series methods and, in general, takes the following form:

$$\mathfrak{L}(y_i, \hat{y}_i) = [R^{-1}\sum_{i=1}^{R} |e_i|^2]^{\frac{1}{2}}, \tag{8}$$

with y_i and $\hat{y}_i$ denoting the observed values and the predictions respectively, e their difference and R the sample size.

b) The extrapolation methods

Empirical evidences and the nature of the problem at hand have been led to discarding a pure standard regression scheme to make inferences on the bivariate vector $\mathcal{L}(\cdot)$ and $\delta(\cdot)$. In fact, the stochastic variability in the data plus the inevitable noise components embedded in the system make difficult to find a solid – statistically significant – relation between the two variables. In addition, as it is well known, being simple regression schemes not designed to take into account the correlation structures embedded in the data, memory information would be lost. This is not a negligible hurdle, as we want our estimations to be affected by the entire process' dynamic and possibly to take in greater account the most recent observations. However, in general modeling past data would require a "not small" number of observations available, especially in consideration of the fact that the proposed method uses block of data of length $\bar{\mathcal{H}}$. In order to satisfy these conditions, two different approaches have been considered, i.e. a polynomial regression ($POLY$) and a double exponential smoothing model (DES). An equation of the type $POLY$ tries to model the functional relationship between two variables by employing basis functions of the type $g(x) \in \mathbb{R}^{d_g}$, e.g. $[(1,x)] \xrightarrow{g} [1, x_i, x_i^2, \ldots x_i^d]$. Its general expression, being y and x respectively the independent and the dependent variable, takes the form $\mathbb{E}[y] = \beta_0 + \beta_1 x + \cdots + \beta_d x^d$, which in matrix forms becomes $\mathbf{y} = \mathbf{X}a + \mathbf{e}$. In this framework, the problem is in general formalized by considering a model of the form $y_i = a_0 P_0(x_i) + a_1 P_1(x_i) + \cdots a_d P_d(x_i)\varepsilon_i, \quad i = 1,\ldots,n$ which is to to be fitted. Notice that the estimation of the term a is done by ordinary least square, i.e. $a = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ which, for a_j take the form $[\hat{a}_j = \frac{\sum_{i=1}^n P_j(x_j)y_j}{\sum_{i=1}^n P_j^2(x_i)}; \quad j = 1,2,\ldots,d]$, whose variance is $V(\hat{a}_j) = \frac{\sigma^2}{\sum_{i=1}^n [P_j(x_i)]^2}$, being the

Ref

9. James MW Brownjohn. Structural health monitoring of civil infrastructure. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1851):589–622, 2007.

generic term $P_r(x_i)$ the $r^{th} order$ orthogonal polynomial. This type of regression scheme has been considered here as it might allow meaningful interpretations of the extrapolation mechanism and can work satisfactorily with a (reasonably) small set of data. In addition, its estimation is in general easy given both the availability of fast and reliable routines and by design: in fact, due to the orthogonality of the polynomials involved, no recomputation of $(X^TX)^{-1}$ or of any other $a_j\,(j \neq k+1)$ is required, so that higher orders polynomial can be introduced at s small cost into the model, e.g. to attempt estimations on a trial and error basis. In this regard, it should be emphasized how the procedure can be easily iterated until a satisfactorily fitting is found. Finally, being technically a special case of multiple linear regression, $POLY$ shares with it the whole, well known, theoretical framework. However, its outcomes can be affected by the non-local nature of the polynomial basis functions, so that the fitted (as well as the extrapolated) values, depend on all the data set, regardless the location in time of the single observations. For the problem at hand, $POLY$ has been employed to model the non-linear relationship $\mathcal{L}(\cdot)$–time and $\delta(\cdot)$–time and the polynomial degree $d = 3$ seemed to yield acceptable predictions. While in $POLY$ the past observations are processed being assigned equal weights, in the second model considered (DES) more recent observations are given higher weights than the older ones, so that the forecast is generated accordingly. In particular, DES is generally represented by the following set of equations:

1. $C_t = \alpha y_t + (1-\alpha)(C_{t-1} + T_{t-1})$
2. $T_t = \beta(C_t - C_{t-1}) + (1-\beta)T_{T-1})$
3. $F_{T+1} = C_t + T_t,$

being: β = trend-smoothing constant, C_t = smoothed constant-process value for period t, T_t = smoothed trend value for period t, F_{t+1}= forecast value for period $t+1$. The use of such an approach is justified by the fact that – as expected – in the empirical experiment always linear memory structures have been always found in the δ and $\mathcal{L}$ sequences. Regarding the parameters estimation procedure, it is based on the minimization of the in sample Mean Square Error. However, a drawback of the DES approach is that in our simulations it has proved to yield more unstable predictions with smaller sample sizes than POLY.

c) The algorithm

Without loss of generality, in what follows it is assumed that:

1. Assumptions:
   (a) $h \equiv H = 1,\ \ (h, H) \in \mathbb{Z}^+$;
   (b) $\mathcal{H} \equiv \bar{\mathcal{H}} \geq 2,\ \ (\mathcal{H}, \bar{\mathcal{H}}) \in \mathbb{Z}^+$ (in the empirical experiment it will be set to 4);
   (c) $\frac{T}{\mathcal{H}} = k,\ \ k \in \mathbb{Z}^+$;
2. Time Series Segmentation:
   (a) the training set $\mathcal{X}_t$, with length N_{tr} , is defined ;
   (b) the test set $\mathcal{Y}_t$ with length N_{ts} , is defined ;
3. Forecast Generation:
   (a) a maximum ARIMA order $(p_0, d_0, q_0, P_0, D_0, Q_0)$, likely to encompass the true model order, is arbitrarily chosen;

(b) optimal MAICE–wise (5) ARIMA model is fitted to the time series at hand (1) conditioned to Tr, i.e. $M^{\bullet}|\mathcal{I}_{Tr} \equiv M^{\circ}|\mathcal{I}_{Tr}$;

(c) ITERATE (3b) $[Nts-1]$ times, i.e. every $x_t \forall t = x_{Nts}, \ldots, x_T$ s.t. the OSH predicted values are stored in the vector conditioned to the last available datum is generated, i.e.

$$\mathbf{y}^{\bullet} \equiv (y^{\bullet}_{(Tr+1)}|\mathcal{I}_{Tr}), \quad (y^{\bullet}_{(Tr+2)}|\mathcal{I}_{Tr+1}), \quad \ldots, \quad (y^{\bullet}_{(Ts)}|\mathcal{I}_{Ts-1}); \tag{9}$$

(d) ITERATE (3b) k times, i.e. every $\bar{\mathcal{H}}$ observations s.t. the OSH predicted values vector y° conditioned to the model fixed every $\bar{\mathcal{H}}$ observations, is generated, i.e.

$$\mathbf{y}^{\circ} \equiv (y^{\circ}_{(Tr+1)}|\mathcal{I}_{Tr}), (y^{\circ}_{(Tr+2)}|\mathcal{I}_{Tr}), \ldots, (y^{\circ}_{(Tr+\bar{\mathcal{H}}-1)}|\mathcal{I}_{Tr}),$$

$$(y^{\circ}_{(Tr+\bar{\mathcal{H}})}|\mathcal{I}_{Tr+\bar{\mathcal{H}}-1}), \ldots, (y^{\circ}_{(Ts-k\bar{\mathcal{H}})}|\mathcal{I}_{Ts-k\bar{\mathcal{H}}-1}), \ldots, (y^{\circ}_{Ts}|\mathcal{I}_{Ts-k\bar{\mathcal{H}}-1}); \tag{10}$$

4. Distance and Loss of Fitting Prediction

(a) The distance measure is sequentially computed on window of length $(\bar{\mathcal{H}})$ of y° and $y^{\bullet}$, i.e. $\delta(y^{\circ}, y^{\bullet})_{Tr+a\bar{\mathcal{H}}}; \quad a = 1,2, \ldots k$;

(b) The loss function is sequentially computed on window of length $(\bar{\mathcal{L}})$ of y° and $y^{\bullet}$, i.e. $\mathcal{L}(y^{\circ}, y^{\bullet})_{Tr+a\bar{\mathcal{H}}}; \quad a = 1,2, \ldots k$;

(c) Standard polynomial regression–based extrapolation scheme is applied to both the functions $\mathcal{L}(\cdot)$ and $\delta(\cdot)$ for the $Nts + \bar{\mathcal{H}}$ period i.e. $\hat{\mathcal{L}}(\bar{\mathcal{H}}) = \mathbf{P}[\mathcal{L}(y^{\circ}, y^{\bullet})]_{Nts+1,\ldots,Nts+\bar{\mathcal{H}}}$ and $\hat{\delta}(\bar{\mathcal{H}}) = \mathbf{P}[\delta(y^{\circ}, y^{\bullet})]_{Nts+1,\ldots,Nts+\bar{\mathcal{H}}}$;

(d) The related expected values are taken, i.e. $\tilde{\mathcal{L}} = \mathbf{E}[\mathcal{L}(y^{\circ}, y^{\bullet})]_{Nts+1,\ldots,Nts+\bar{\mathcal{H}}}$ and $\tilde{\delta} = \mathbf{E}[\delta(y^{\circ}, y^{\bullet})]_{Nts+1,\ldots,Nts+\bar{\mathcal{H}}}$, i.e. $\hat{\mathcal{L}}(\bar{\mathcal{H}}) = \frac{1}{\bar{\mathcal{H}}}\sum_{j=1}^{\bar{\mathcal{H}}} \mathbf{P}[\mathcal{L}(y^{\circ}, y^{\bullet})]_{Nts+j}$ and $\hat{\delta}(\bar{\mathcal{H}}) = \frac{1}{\bar{\mathcal{H}}}\sum_{j=1}^{\bar{\mathcal{H}}} \mathbf{P}[\mathcal{L}(y^{\circ}, y^{\bullet})]_{Nts+j}$.

d) Empirical Experiment

This section is devoted to the empirical experiment which has been designed and caried outin order to test the validity of the proposed procedure. It consists of two parts: a Monte Carlo experiment, based on computer generated time series and an analysis of four real-life time series, two of the type Macroeconomic and two related to tourism variables. Regarding the Monte Carlo experiment, four different DGPs – whose parametrization is given in Tab.1 along with the codification used for brevity and reported in the column labeled "DGP" – have been employed to generate 1000 realizations (250 realizations for each model), with sample size $t = 300$. The main reason behind the choice of series showing such a limited sample size is that instabilities in the ARIMA parameters are more likely to occur under small sample sizes and therefore greater uncertainty is expected in terms of both $\delta(y^{\circ}, y^{\bullet})$ and $\mathcal{L}(y^{\circ}, y^{\bullet})$. In addition, such a situation is common in economic time series but also in all the cases where only a small set of past data is subjected to investigation, e.g. due to computational reasons. In order to mimic reality, realizations of DGP 1–4 are corrupted with short bursts of noise (iid shocks) in the form of outliers of the type additive (AO). Such a sequence of isolated spikes have been introduced to represent those noticeable

departures – consistently found across the empirical experiment – that sporadically might take place in the series $\delta(\cdot)$ and $\mathcal{L}(\cdot)$, as a result of the effect of sudden changes on the models. To do so, $\mathbf{O} = 3$ Additive Outliers have been embedded in the test set y_t, so that the resulting set up can be formalized as follows:

$$y_t^* = \sum_{j=1}^{\mathbf{O}} \xi_j(B)\gamma_j I_t^{(\psi_j)} + y_t \,, \tag{11}$$

being y_t^* the stretch of data corrupted by the outliers, y_t its outlier-free, unobservable, counterpart (3) and γ_j represents the outlier's impact at ψ_j and I_t is a switching variable allowing the system to (not) include the outlier in $t = \psi_j$ when $I = 1(0)$. Training and Test sets' sample sizes have been set respectively at nTs =180 and nTr =120 whereas the outliers have been embedded in the test set at observations $t = Nts/4, Nts/3, Nts/2$. Their values have been kept fixed and set to 6 σ^2, being $\sigma^2 = 1 \ \forall \ DGPs$.

Table 1: Parametrization of the simulated DGPs

DGP number	ARIMA order	ϕ	ϕ
DGP1	(0,1,1)	–	-.6
DGP2	(1,1,2)	-.65	.6; -.45
DGP3	(2,0,1)	.7; -.5	-.5
DGP4	(1,0,2)	-.6;	.5; -.4

Table 2: Actual vs predicted distances and loss functions in the simulated time series case: percentage difference

DGP	$\widehat{\mathfrak{L}}_{\%}(y^{\bullet}, y^{\circ})$		$\widehat{\delta^{CI}}_{\%}(y^{\bullet}, y^{\circ})$		$\widehat{\delta^{NIPER}}_{\%}(y^{\bullet}, y^{\circ})$		Ntr	Nts
	Poly	DES	Poly	DES	Poly	DES		
DGP1	10.2	11.6	10.5	11.9	11.1	12.2	220	80
DGP2	12.4	14.3	10.4	11.2	12.6	10.9	220	80
DGP3	8.0	7.2	9.4	4.7	8	6.3	100	200
DGP4	6.2	5.8	9.5	5.4	8.2	7.5	100	200

Table 3: Actual vs predicted distances and loss functions in the real time series case: percentage difference

DGP	$\widehat{\mathfrak{L}}_{\%}(y^{\bullet}, y^{\circ})$		$\widehat{\delta^{CI}}_{\%}(y^{\bullet}, y^{\circ})$		$\widehat{\delta^{NIPER}}_{\%}(y^{\bullet}, y^{\circ})$		Ntr	Nts
	Poly	DES	Poly	DES	Poly	DES		
DGP1	15.3	18.9	16.5	19.9	22.1	19.6	300	56
DGP2	15.4	10.1	14.4	12.5	22.6	20.1	300	120
DGP3	8.9	9.5	15.6	7.3	10.7	10.7	219	120
DGP4	9.7	8.6	10.5	6.8	8.7	8.7	219	120

Table 4: Real time series employed in the empirical section: sources and main details

Code	Variable	Source	Seas	Units	Data range (Number of obs)
X1	Housing: mortgage interest payments	Data Set MM23 (San Louis Fed)	No	Index, base 1987 = 100	1987-02 to 2016-09(356)

X2	Consumer Price Index for All Urban Consumers: All Items	US. Bureau of Labor Statistics	No	Index, base 1984 = 100	1981-10 to 2016-09(420)
X3	OS visits to UK: Earnings: Θ Millions	U.K. Office for National Statistics	YES	Θ Millions	1980-01 to 2016-07 (439)
X4	OS visits to UK: All visits	U.K. Office for National Statistics	YES	Thousands of visitors	1980-01 to 2016-07 (439)

Notes

Regarding the second part of the experiment, in Table 4, the four time series employed in the empirical study are detailed along with their conventional name, in the sequel adopted for brevity, stored in the column labeled "Code". Series X1 – X2 are of the type macroeconomic, whereas the remaining ones refer to tourism-related variables. All the time series are characterized by a limited sample sizes (not too far from the one of the computer generated time series), the presence of outliers – e.g. of the type additive, as clearly noticeable in the series X4 (May 2013 2006) and X3 (July 2007) – and, to a different extent, non stationary behaviors. All the series have not been adjusted for seasonality nor corrected for outliers. Finally, the variable "seas" in Table 4 indicates the presence of a significant seasonal component in the series, which has been properly captured by the seasonal parameters of the seasonal version of the ARIMA model.

i. *Experiment's outcomes*

Regarding the Monte Carlo experiment, the mean values of the loss function and the distance metrics have been computed over each set (250 series), i.e. $\hat{\delta}_{mc}(\overline{\mathcal{H}}) = \frac{1}{250}\sum_{1=1}^{250}\uparrow(\sum_{j=1}^{4}\uparrow\hat{\delta}(\mathcal{H}_j))$ and $\hat{\mathcal{L}}_{mc}(\overline{\mathcal{H}}) = \frac{1}{250}\sum_{1=1}^{250}\uparrow(\sum_{j=1}^{4}\uparrow\hat{\mathcal{L}}(\mathcal{H}_j))$, with $\overline{\mathcal{H}} = 4$ and the subscript "mc" standing for Monte Carlo. In Tables 2 and 3 – where the results of the empirical experiment are reported – the following two indicators are employed to evaluate the usefulness of the proposed procedure, i.e. the Loss function discrepancy percentage change and the the Distance Discrepancy percentage, respectively defined as follows: $\hat{\mathfrak{L}}_{\%}(y^{\bullet}, y^{\circ}) = 100\frac{\hat{\mathfrak{L}}_{mc}(\overline{\mathcal{H}}) - \mathfrak{L}_{mc}(\overline{\mathcal{H}})}{\hat{\mathfrak{L}}_{mc}(\overline{\mathcal{H}})}$ and $\hat{\delta}_{\%}(y^{\bullet}, y^{\circ}) = 100\frac{\hat{\delta}_{mc}(\overline{\mathcal{H}}) - \delta_{mc}(\overline{\mathcal{H}})}{\hat{\delta}_{mc}(\overline{\mathcal{H}})}$.

The results obtained indicate the interesting prediction capabilities provided by the proposed procedure, which can be considered adequate to gain valuable insights on the discrepancies resulting from the use of a fixed ARIMA model instead of its updated version. With both artificially generated and real time series, the best performances are obtained – under the condition of a test set of "sufficient" length – by using the exponential smoothing extrapolation technique in conjunction with a distance metric of the type CI. On the other hand, less impressive outcomes are obtained with small test sets. In this case the polynomial regression has yielded slightly better outcomes than the exponential smoothing scheme. However, even for small values of Nts, the approach still seems to provide useful information, especially in terms of expected loss function, where the percentage difference under polynomial regression recorded is around 10.2% and 11.6% in the case of the artificial time series DGP1 and DGP2 respectively and slightly higher (15.3%) for the real time series TS1. For larger test sets the DES extrapolation technique does a better job than the regression–based technique: the recorded value for

$\mathcal{L}_{\%}(y^{\bullet}, y^{\circ})$ and $\delta_{\%}^{CI}(y^{\bullet}, y^{\circ})$ is always less than approx 10% and 12.5% respectively. In the set of the real time series, the best performances have been obtained in the case of TS4, where an error of 6.8% and 8.6% have been recorded for the distance CI and the RMSE values respectively, computed via DES equations. Throughout the empirical experiment, the values recorded for δ^{NPER} has been consistently less remarkable results. A possible explanation is related to the sensitivity of the periodogram towards aberrant observations, so that bias components might have been introduced into its estimation as a result.

III. Conclusion

In this paper, it has been illustrated a procedure for the prediction of the lack of fit and the distance between the outcomes of two models, when one of them is re–estimated at the highest possible frequency (i.e. the sample frequency of the time series under investigation) and the other one is left unchanged for a certain span of time. This technique has been presented using time series models belonging to the class ARIMA, however, such a conditions can be easily relaxed and basically left to be decided on a case-by-case basis. All the simulations have been carried out having in mind a short span of time, set to 4, between two updating processes and the results turned out to be encouraging. In particular, consistency in terms of empirical outcomes has been found across the statistical tools employed and the time series used. Finally, out of the sets of the available extrapolation techniques and distances measures, only two pair of them have been here considered, so that future directions will include the analysis of a larger portfolio of these tools.

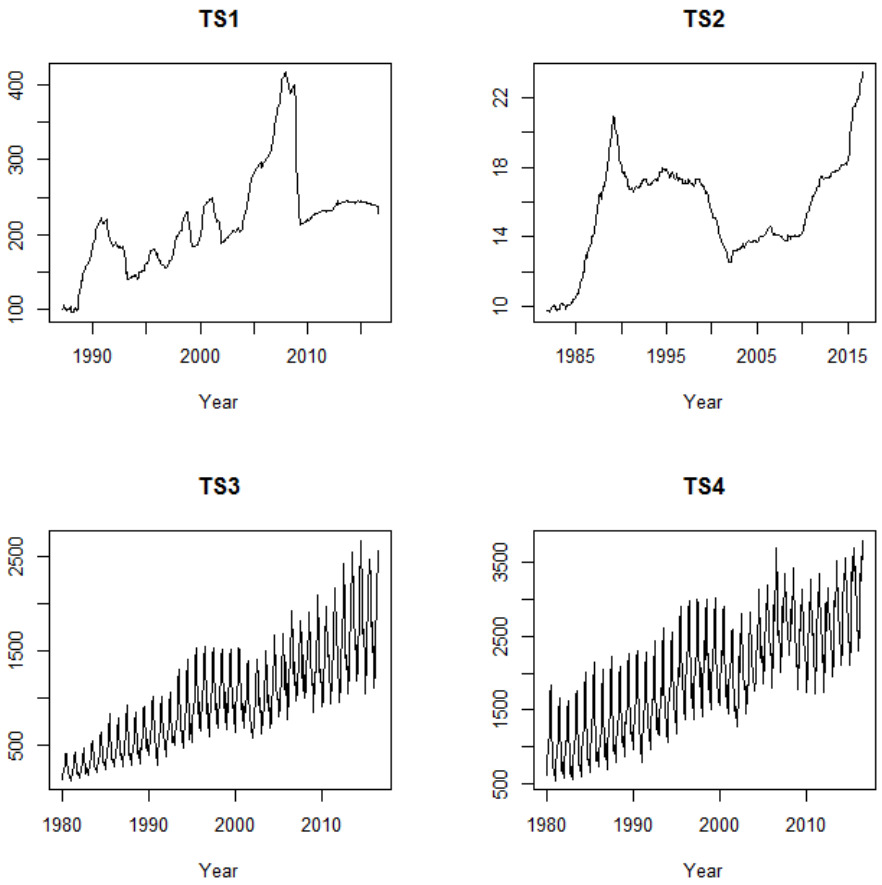

Figure 1: Actual Time Series

References Références Referencias

1. David F Findley, Brian Monsell, Mark Otto, William Bell, and Marian Pugh. Towards x12 arima. Technical report, mimeo, Bureau of the Census, 1992.
2. Victor Gomez and Agustin Maravall. Programs tramo (time series regression with arima noise, missing observations, and outliers) and seats (signal extraction in arima time series). instructions for the user. *Documento de Trabajo*, 9628, 1996.
3. David F Findley and Catherine C Hood. X-12-arima and its application to some italian indicator series. *Seasonal Adjustment Procedures–Experiences and Perspectives*, pages 231–251, 1999.
4. Christophe Planas and Raoul Depoutot. Controlling revisions in arima-model-based seasonal adjustment. *Journal of Time Series Analysis*, 23(2):193–213, 2002.
5. G Huyot, Kim Chiu, John Higginson, and Nazira Gait. Analysis of revisions in the seasonal adjustment of data using x-11-arima model-based filters. *International Journal of Forecasting*, 2(2):217–229, 1986.
6. Jens Mehrhoff and Deutsche Bundesbank. Sources of revisions of seasonally adjusted real time data. In *5th EUROSTAT Colloquium on Modern Tools for Business Cycle Analysis, Luxembourg*, 2008.
7. Scott W Doebling, Charles R Farrar, Michael B Prime, and Daniel W Shevitz. Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Technical report, Los Alamos National Lab., NM (United States), 1996.
8. Hoon Sohn, Charles R Farrar, Francois M Hemez, Devin D Shunk, Daniel W Stinemates, Brett R Nadler, and Jerry J Czarnecki. A review of structural health monitoring literature: 1996–2001. *Los Alamos National Laboratory, USA*, 2003.
9. James MW Brownjohn. Structural health monitoring of civil infrastructure. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1851):589–622, 2007.
10. Wei-Xin Ren and Guido De Roeck. Structural damage identification using modal data. ii: Test verification. *Journal of Structural Engineering*, 128(1):96–104, 2002.
11. Michael I Friswell. Damage identification using inverse methods. In *Dynamic methods for damage detection in structures*, pages 13–66. Springer, 2008.
12. George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
13. Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
14. Hirotugu Akaike. Akaike information criterion. In *International Encyclopedia of Statistical Science*, pages 25–25. Springer, 2011.
15. Gustavo EAPA Batista, Xiaoyue Wang, and Eamonn J Keogh. A complexity-invariant distance measure for time series. In *SDM*, volume 11, pages 699–710. SIAM, 2011.
16. Gustavo EAPA Batista, Eamonn J Keogh, Oben Moses Tataw, and Vincius MA de Souza. Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3):634–669, 2014.
17. David Casado de Lucas. Classification techniques for time series and functional data. 2010.

Notes

This page is intentionally left blank