
The intelligent forecasting model of time series

Sonja Pravilović^{1, 2} Annalisa Appice²

¹Montenegro Business School, "Mediterranean" University, Podgorica, Montenegro

²Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari, Italy

Email address:

sonja.pravilovic@uniba.it (S. Pravilovic), annalisa.appice@uniba.it (A. Appice)

To cite this article:

Sonja Pravilović Annalisa Appice. The Intelligent Forecasting Model of Time Series. *Automation, Control and Intelligent Systems*. Vol. 1, No. 4, 2013, pp. 90-98. doi: 10.11648/j.acis.20130104.12

Abstract: Automatic forecasts of univariate time series are largely demanded in business and science. In this paper, we investigate the forecasting task for geo-referenced time series. We take into account the temporal and spatial dimension of time series to get accurate forecasting of future data. We describe two algorithms for forecasting which ARIMA models. The first is designed for seasonal data and based on the decomposition of the time series in seasons (temporal lags). The ARIMA model is jointly optimized on the temporal lags. The second is designed for geo-referenced data and based on the evaluation of a time series in a neighborhood (spatial lags). The ARIMA model is jointly optimized on the spatial lags. Experiments with several time series data investigate the effectiveness of these temporal- and spatial- aware ARIMA models with respect to traditional one.

Keywords: Time Series Analysis, Arima, Auto. Arima, Lag. Arima

1. Introduction

In recent years, globalization has significantly accelerated the communication and exchange of experience, but has also increased the amount of data collected as a result of monitoring the spread of economic, social, environmental, atmospheric phenomena. In such circumstances, it is necessary a useful tool for analyzing data that represents the behavior of these phenomena and drawing useful knowledge from these data to predict their future behavior.

Accurate forecasts for the phenomenon behavior can make anticipation of the actions (for example, the prediction of wind speed in a region allows us to define the better strategy to maximize profit in the energy market).[16]

In the last two decades, several models of (complex) time series dynamics have been investigated in statistical analysis.[3] Forecasting algorithms must determine an appropriate time series model, estimate the parameters and compute the forecasts. They must be robust to unusual time series patterns, and applicable to large numbers of series without user intervention. The most popular automatic forecasting algorithms are based on the ARIMA model, which optimizes the parameters of the model for a single univariate time series.[6] In this way, multiple univariate time series, such that, geo-referenced time series, which are

generated by several sensors of a network, are modeled separately. This naive approach neglects the spatial component of time series, which refer points placed at specific spatial locations. When analyzing geo-referenced data, we frequently the phenomenon of spatial autocorrelation.[18]

Spatial autocorrelation is the correlation among the values of a single variable (i.e., object property) strictly attributable to the relatively close position of objects on a two-dimensional surface, introducing a deviation from the independent observations assumption of classical statistics. Intuitively, it is a property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for pairs of observations at randomly selected locations (Moran 1950).[20] Positive autocorrelation is common in spatial phenomena (Goodchild 1986).[18] Spatial positive autocorrelation occurs when the values of a given property are highly uniform among spatial objects in close proximity, i.e., in the same neighborhood. In geography, spatial autocorrelation is justified by Tobler's first law (Tobler 1970)[17], according to which "Everything is related to everything else, but near things are more related than distant things".[17] This means that by picturing the spatial variation of some observed variables in a map, we may observe regions where the distribution of values is

smoothly continuous, with some boundaries possibly marked by sharp discontinuities. This suggest that a forecasting model is smoothly continuous in neighborhood and inappropriate treatment of data with spatial dependencies, where spatial autocorrelation is ignored, can obfuscate important insights in the models.

In this paper we formulate an inference procedure that allows us to and to obtain a robust and widely applicable automatic forecasting algorithm which optimizes the traditional ARIMA model by "jointly" estimating forecasting parameters for several time series lags.[12] These lags can be consecutive seasons of a single time series or multiple time series with spatial dependence. This algorithm has been implemented in the forecast package for R. For each (temporal- or spatial-aware) lag we apply all models that are appropriate, optimizing the parameters of the model in each case and selecting the best model according to the AIC. The point forecasts can be produced by using the best model (with optimized parameters) for as many steps ahead as required.

The paper is organized as follows. In the next Section we revise basic concepts and background of this work. In Section 3, we describe the algorithm *lag.arima* in its temporal and spatial formulation. In Section 4, we report results of an empirical evaluation on real-world times. Finally, some conclusions are drawn in Section 5.

2. Background

The Box-Jenkins approach is one of the most popular and powerful forecasting technique. It is an ARIMA model [2], which is a generalisation of an ARMA model [12].

The autoregressive-moving-average (ARMA) model describes a (weakly) stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average.[21]

$$X_t = c + \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (1)$$

The auto-regressive model of order p . refers to the moving average model of order p .

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t \quad (2)$$

where φ_i are parameters, c is a constant, and the random variable ε_t is white noise.

MA(q) refers to the moving average model of order q .

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (3)$$

φ_i are parameters of the model, μ is the expectation X_t (often assumed to equal 0), and the $\varepsilon_t, \varepsilon_{t-1}, \dots$ are white noise error terms.

Gershenfeld and Weigand [5] indicate that the selection of ARMA order (p, q) models is not simple. ARMA models in general can, after choosing p and q , be fitted by least squares regression to find the values of the parameters which minimize the error term. It is generally considered good practice to find the smallest values of p and q which provide an acceptable fit to the data [3] recommend using AICc for finding p and q .

Time series were mainly studied under a deterministic aspects, until in 1927 Yule [15] introduced the notion of stochasticity. According to him, every approach to time series can be regarded as the realization of a stochastic process. This idea of stochastic process launched a different number of time series methods, varying in parameter estimation, identification, forecasting and checking method.

Box and Jenkins in their publication Time Series Analysis: Forecasting and Control [3] integrated the existing knowledge and made a breakthrough in the area creating a coherent and versatile approach identifying the three stage iterative cycle for time series: identification, estimation and checking diagnostic. The autoregressive integrated moving average (ARIMA) models by the evolution of computers made the use more popular and applicable in many scientific fields.

Auto.arima function [9] conducts a search over possible model within the order constraints provided and returns the best ARIMA model according to either AIC, AICc or BIC value. Non-stepwise selection can be slow, especially for seasonal data. Stepwise algorithm outlined in Hyndman and Khandakar [9] except that the default method for selecting seasonal differences is now the OCSB test rather than the Canova-Hansen test. *Auto.arima* function that is within the library (forecast) in the R programming language has the task to define the parameters p, d and q having a minimum AIC criterion. [9].

The combination of forecasts is a widely investigated issue in the statistical field. Many researchers have recognized the value of combining forecasts produced by various techniques as a means of reducing forecast error [4]. Armstrong's meta analysis [14] marked that combining of different model can be more useful for short range forecasting, where random errors also can be more significant. As because these errors are off setting, a combined prediction methods should reduce them. Many combining forecasts techniques have been introduced over the past years, varying in level of success and complexity.

Timmerman argue that simple combinations that ignore the correlation between the predicted errors often dominate over refined combination schemes aimed at assessing theoretically optimal combination weights. The use of a simple average has proven to do as well as more sophisticated approaches. Nevertheless, there are situations where one method is more accurate than another. If such cases can be identified in advance, simple averages would be inefficient [14].

Applying Box-Jenkins methodology optimization parameter there are available various packages to find the

right parameters for the ARIMA model. In this paper, we used R functions included in the standard stats package, which includes *arima* function documented in 'ARIMA Modelling of Time Series' [7,8]. The *ARIMA(p,d,q)* function also includes seasonal factors, an intercept term, and exogenous variables called 'external regressors'. The 'forecast' package with the *auto.arima()* function in R automatically select the best ARIMA model for a given time series.

The forecast package for the R system for statistical computing (R Development Core Team 2008) is part of the forecasting bundle [7,8,9,10,11] is available from the Comprehensive R Archive Network at <http://CRAN.R-project.org/package=forecasting>. Version 2.15 of the package was used for this paper. The forecast package contains functions for univariate forecasting and implements automatic forecasting using exponential smoothing, ARIMA models, the Theta method (Assimakopoulos and Nikolopoulos 2000), cubic splines [10,11], as well as other common forecasting methods.

But in classical statistics, correlation analysis has as a basic assumption independence between the variables. This assumption does not exist in spatial statistics in which the observed values depends with each other. In fact, it is assumed that the observations or measured values in near places are more connected and are very similar. In other words, what the station which measured variables are further, it would measured values more different to each other.

Automatic prediction algorithms must determine the appropriate time-series model to estimate the parameters and calculate the forecast, but the most popular prediction algorithms and techniques only treat space or time dimension, a rarely time and space simultaneously.

The idea is to use these methods of predicting univariate and multivariate time series, which however do not take into account the spatial dependence to get more information needed for better prediction.

This article presents a new methodology that provides short-term forecasts of natural phenomena first form and then dividing time series in specific lags and exploring the unique and best predicting parameters model valid for each lag (for all sensors in closely area) of the time series. This is because, our intention is experimentation and applying algorithms that use the methodology and techniques of time series predictions explaining the fact that the statistical properties of the recorded data (prediction in time series behaviour) can be applicable not only to the dimension of time, but also space. In doing so, the goal of this empirical study is to analyze accuracy and efficiency of our proposed algorithm, dividing the time series in lags and calculating the mean absolute prediction errors and the mean of residuals for all lags, instead of calling *auto.arima* which gives only one group of best prediction parameters for all-time series. By comparing the result of prediction using those parameters and the result of *auto.arima* for all-time series we demonstrate that our model is more accurate and

provides a detailed analysis of future behaviour of time series. Our research aims to take advantage of the methodology of univariate time series forecasting techniques such as *arima*, *auto.arima*, and others in predicting measurable values of sensors that are distributed in spatial dimension by grouping in zones, and predict the measurable values of the sensors founded in their neighbourhood.

3. Algorithm

It is possible to take advantage of the methodology of time series forecasting techniques such as *arima*, *auto.arima*, and others in predicting measurable values (for example sensors) that are distributed in spatial dimension to predict measurable values in the neighbourhood.

The main task of this paper was to show that dividing time series in the time intervals of specified length (lags) and then finding the best and unique model parameters valid for all lags needed to make predictions of future values provides a big advantage. By comparing the result of prediction using those parameters and the result of *auto.arima* for all-time series we demonstrate that our model is more accurate and provides a detailed analysis of future values of time series.

The first idea to solve problem of forecasting of such formed time series with application of *auto.arima* function was dividing time series in the lags, and then for each lag call *auto.arima* function that finds the best parameters. But, dividing the time series at lags and then using the function *auto.arima* for each lags occur two kinds of problems.

1. For each lag, *auto.arima* finds the best parameters p , d and q , regardless of whether that part of the interval of lag is stationary or not. Acceptance of the parameters that give the optimal solution in one lag may prove to be a complete failure in the other (in case that the part of the series in the next lag is not stationary or has completely different flow). Thus, the obtained model and parameters p , d and q , that are commonly used for each lag gives the results, but the solution is not general, since it can simultaneously apply more equal number of times or obtained solution is not applicable for each lag of time series. Some lags of time series can also be non-stationary and the parameters p , d and q , for all lags in general do not give a meaningful result).
2. Inefficient multi calling function *auto.arima* and different number of parameters for each lag.

Thus, our algorithm has the task to applied modified *auto.arima* function that defines the length of the time interval (after that period the time series repeats or has a very similar behavior). This is because the time series consists of data obtained from the sensors that are located closely and measured very similar values of the phenomenon in the same period of time.

Our algorithm examine the unique parameters that are valid for each lag of whole time series and looking for

those parameters for which average AIC is minimal for all lag, since they allow to obtain unique and best predicting parameter of future behavior of the time series. The idea is that the algorithm which uses the methodology of time series forecasting techniques, while explaining the fact that the statistical properties of the recorded data can be applicable not only to the dimension of time, but also space.

Thus, as the input data set there are more short time series of measured data on spatially distributed sensors, which are located closely and is formed only one time series for the same period of time.

Our algorithm apply modified *auto.arima* model for the entire time series, which is divided into lags (values measured on the sensor in near neighborhoods) and find a unique and best model that gives the minimum average value of AIC's which passed all lags of formed time series. In the cases that are possible more than one and same model (parameters p , d and q) to apply to all lags (time intervals) will be selected that one which for the entire time series, for each lag, chooses the one that give the minimal average AIC.

In order to conduct a more precise forecasting of longer time series as at specific time intervals show repeats flow, in this paper, we propose a new prediction algorithm as follows:

First, we display the entire time series and observed length of time interval after which the time series behaves about the same or similar way.

Second, we call modified function *auto.arima* within the R software package, which we called *lag.arima* whose task is to identify the ARIMA model with the same parameters p , d , q , P , D , Q , which are valid for each lag of time series. In the event that there are multiple models that are valid for each lag, *lag.arima* will selected as the optimal model and its parameters one whose average AIC is minimal for all lag.

As we analyzed at the time series in the following figure 1, we can see the characteristic movement repeated similar values 12 times. Since the timestamp series contains 4032 values related to the half-hourly demands for electricity, when divided by 12, we can get the length of a period or 336 values.

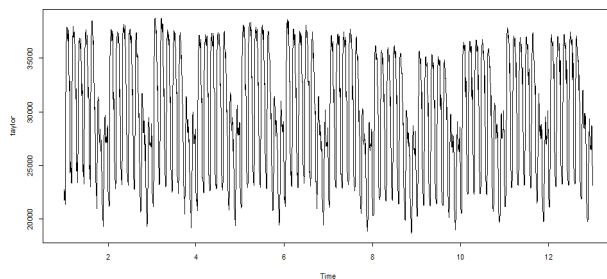


Figure 1. The time series 'Taylor'

Having established the length of a lag, we called the modified functions *auto.arima*, which we called *lag.arima*.

The time series is formed in such way to take elements of the sensor that is located in the center, and below all other values of sensors located closely (immediate neighborhood) to the central sensor for a defined period of time (example 24 hours), then form input data set, i.e. the time series W as follows:

| Sensor ID | Wind[t1] | Wind[t2] | ... | Wind[24] | ... |
|-----------|----------|----------|-----|----------|-----|
| Sensor 1 | w1[2] | w1[2] | ... | w1[24] | ... |
| Sensor 2 | w2[1] | w2[2] | ... | w2[24] | ... |
| ... | ... | ... | ... | ... | ... |
| Sensor k | wk[1] | wk[2] | ... | wk[24] | ... |

The first period (lag) of time series can be represented by figure 2.

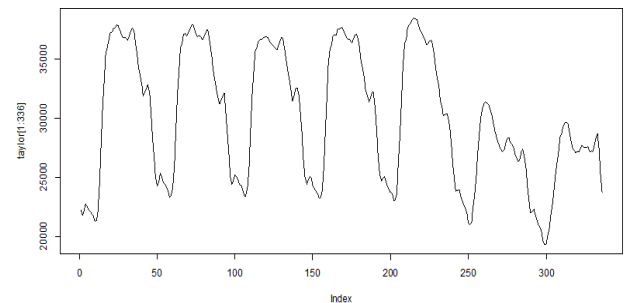


Figure 2. The first lag of time series

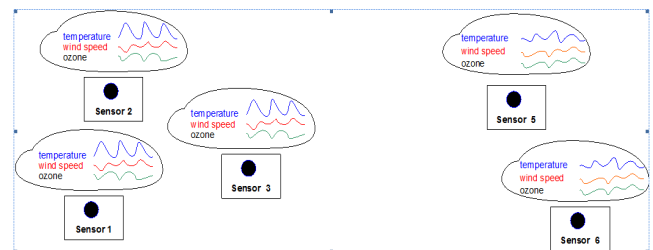


Figure 3. The spatial distribution of the sensor with the time series of measured parameters

Time series: $w1[1]$, $w1[2]$, ..., $w1[24]$, $w2[1]$, $w2[2]$, ..., $w1[24]$, and $\text{lag}=24$.

As has already been explained at the beginning of this paper, *auto.arima* function of R software package is based on the minimum value of AIC calculating optimal ARIMA model parameters p, d, q, P, D, Q for the entire time series.

Our algorithm sets all possible parameter values (as it does *auto.arima* for the whole time series choosing as optimal one with minimum of AIC). Only those parameters of model that are valid for each period are preserved in the matrix results which besides the parameters p , d , q , P , D , Q , contains also the average value of AIC for each period of the time series. So, while the output values of *auto.arima* function are $ar1$, $ar2$, ..., $ma1$, $ma2$, ... intercept, the minimum AIC, BIC, estimated σ^2 , log likelihood, the output values of our algorithm are the various models parameters p , d , q , P , D , Q , constant and average value of

AIC for all lags and valid for whole time series.

Based on these parameters, and in particular based on the average value of AIC for each model conclusion are the following:

- if there are some period of time series stationary or not;
- which ARIMA model can be applied to the whole, and not on any lag (or time interval) period of the time series;
- evaluate the accuracy of predictions of future values of the time series.

Though our approaches we gave adequate successful forecasts separately to demonstrate accurateness in predictions of future value of time series for a short time period. An application of the ARIMA model would result for predicted values are higher than the actual data and can be explained by the fact that the ARIMA modelling in forecasting is mainly based on the recent historical data.

4. Experimental Results

The *auto.arima()* function in R uses a variation of the Hyndman and Khandakar algorithm which combines unit root tests, minimization of the AICc and MLE to obtain an ARIMA model. The algorithm follows steps as follows:

- The number of differences d is determined using repeated KPSS tests.
- The values of p and q are then chosen by minimizing the AICc after differencing the data d times. Rather than considering every possible combination of p and q , the algorithm uses a stepwise search to traverse the model space.

Auto.arima tries all models for time series (finding one with the smallest AICc) and then selecting the best from the following four. ARIMA(2,d,2), ARIMA(0,d,0), ARIMA(1,d,0), ARIMA(0,d,1). If $d=0$ then the constant c is included; if $d \geq 0$ then the constant c is set to zero.

The best model considered becomes the new model until no lower AICc can be found.

Modeling procedure consists of fitting an ARIMA model to a set of time series data. A useful general approach can be provides following next procedure.

- Plot the data and identify any unusual observations;
- Transform the data using a Box-Cox transformation to stabilize the variance and if necessary;
- If the data are non-stationary take the first differences of the data until are stationary;
- Examine the ACF/PACF to understand if is an AR(p) or MA(q) model appropriate;
- Try chosen model using the AICc to search a better model.
- Check the residuals from chosen model by plotting the ACF of the residuals and making various tests of the residuals. If they do not look like white noise, try another modified model.
- Once the residuals look like white noise then is advisable, calculate forecasts.

In a phase of choosing the order of Box-Jenkins ARMA processes it is necessary automatic procedure using statistically based set of rules. In this sense, the proposed numerous criteria for model selection the most commonly used is AIC criterion.

The Akaike information criterion (AIC) is a measure of the relative goodness of fit of a statistical model. The AIC is grounded in the concept of information entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality. AIC values provide a means for model selection in analysis of time series.

In the general case, the AIC is $AIC_k - 2kn(L)$ where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated forecasting model of time series.

The values p and q in our algorithm are chosen to minimize mean $AIC(p,q)$ for all lags. For evaluation of unknown parameters in time series models commonly are used method of maximum likelihood. On the renewal of the sample (x_1, \dots, x_n) is formed maximum likelihood function.

In this article, we primarily used the ARIMA modelling approach to automatic forecasting and we describe the implementation of this methods in the forecast package, along with other features of the package with some our transformation in the *auto.arima* model.

The forecast package in R language contains the function *Arima()* which is largely a wrapper to the *arima()* function in the package stats. The *Arima()* function makes easier including a drift term when $d + D = 1$. Setting *include.mean=TRUE* in the *arima()* function from the stats package will only work when $d + D = 0$. The facility provides fitting function in ARIMA model to a new data set and the ARIMA models gives a possibilities to one-step forecasting available via *fitted()* function.

Therefore, this paper introduces a methodology for combining the two methods in this special case of forecasting, taking into account the above mentioned facts. The methodology relies on the use of the two predictions of the most suitable ARIMA model measured by the Mean Absolute Error (MAE) and mean of residuals (MR) incorporated into model as actual data in order to construct a new one.

MAE and MR are frequently used measure of the differences between predicted values by a model or an estimator of the actually observed values. The individual differences called residuals present the calculations performed over the data sample used for estimation. There are called prediction errors when computed out of sample. The MAE and MR as the errors in predictions are good measure of accuracy.

Our hypothesis is that the short term forecast of many similar or equal lags produced by the ARIMA models are better than the one produced by *auto.arima* model at this stage of the process, which is validated in the present work. As this forecast is anticipated to be over-optimistic, the predicted values are taken into consideration in order to construct a new model, which will be more effective than

the second one. After the calculation of the new model's parameters, the first two values are recalculated. These forecasts incorporate the effects of the ARIMA and *auto.arima* models which anticipated to be more precise than each approach separately. This procedure uses the simple averages approach, which would not be sufficient in cases of only ARIMA or *auto.arima* models and produce notice able to improve short-term predictions.

The use of weights as a method of combinations is also avoided, as it involves personal judgement regarding their value or evaluation of correlations between forecast errors that can change from period to period.

4.1. Experimental Data

For the experiments, we took three groups of data. The first is a time series of half-hourly electricity demand in England and Wales from Monday 5 June 2000 to Sunday 27 August 2000. Discussed in Taylor [13], and provided by James W. Taylor.

The second one is atmospheric concentrations of CO₂ expressed in parts per million (ppm) reported in the preliminary 1997 SIO manometric mole derived from in situ air samples collected at Mauna Loa Observatory, Hawaii. <ftp://cdiac.esd.ornl.gov/pub/maunaloa-co2/maunaloa.co2>. It is a time series of 468 observations registered monthly from 1959 to 1997. There are missing values for February, March and April of 1964 and they have been obtained by linear interpolation between the values for January and May of 1964. The monthly values have been adjusted to the 15th of each month. Missing values are denoted by -99.99. The 'annual' average is the arithmetic mean of the twelve monthly values. In years with one or two missing monthly values, annual values were calculated by substituting a fit value (4-harmonics with gain factor and spline) for that month and then averaging the twelve monthly values. [21]

The third group of data relates to the Australian monthly gas production in the period 1956-1995.

4.2. Experimental Methodology

The accuracy of each approach of research and applications of prediction model is measured by deviation error from the real value. Our approach is therefore also evaluated on base of two measures of accuracy, mean absolute error (MAE) and mean of residuals (MR) to verify the superiority of our approach versus only *arima* or *auto.arima* model for the entire time series. This article presents a new methodology that provides short-term forecasts of behavior of natural phenomena that divides the time series at specific time intervals (lags), and explores the best model parameters values valid for each lag of the time series.

The goal of forecasting of time series analysis starts from the available data from the past necessary to formulate and evaluate the time-series model and then used it to predict future values of the series. In doing that are used series of

statistical tests and criteria that verifies the validity of the evaluated model.

In this paper, in the analysis and prediction are applied the class of autoregressive moving average model ARIMA (p, d, q). In this class of models is the assumption that the current value (element) of series depends: (1) on the value of previous members of the series, (2) the current value of the random process and (3) the previous value of the random process. In time series with observed effect of the trend, cyclical and seasonal components, the application of these models includes prior removal of that influence. To eliminate the influence of systematic components from the time series is used the operator of differentiation d .

For the identified model, the next step was a recursive procedure of estimation of model parameters or fitting the model. The basic approaches of fitting models are method of nonlinear least squares errors and the method of maximum likelihood.

Once adopted the appropriate model requires estimations of parameters and methodology requires verification of the residuals. If the residuals are random, the model is appropriate. Several tests like the Box-Pierce statistics can be proposed to determine the randomness of the residuals. Otherwise, it is necessary to return to the stage of identifying the model and try to find a better one.

As at any lag the time series can be not stationary, it is necessary differencing of all time series. Stationary time series lag has a constant mean value and variance. Non-stationary at any lag is corrected by corresponding differencing of all values of time series. In this case, the first order of difference and first order of seasonal differences are sufficient to achieve a stationarity of all time series.

The goal is to find such forecasting parameters p , d , q that give the lowest mean values of AIC applicable to each further lag of time series, and on the basis of whose results it is possible to forecast behavior in the coming intervals that show similar or the same behavior, not examining the whole time series always from the beginning, thus giving more accurate predictions.

Table 1-8. MAE for training - testing data set of two models

| co2 (1:1) Training/ Testing set | Training set | Testing set |
|------------------------------------|------------------------------|--------------------------------|
| | x<-co2[1:228] Lag=12 x 19 | x<-co2[229:456] Lag=12 x 19 |
| Forecast arima(2,1,0) | Mae | Mae |
| Abs error of prediction | 0.4373796 | 0.5992418 |
| Mean of residuals | 0.02405235 | 0.03207524 |

| Auto.arima: Training set | | |
|--|--------------|-------------------------|
| x<-co2[1:228] - without division in lags - Training set | | Abs error of prediction |
| Mean of residuals | -0.005477198 | 0.1242049 |
| Arima (2,1,1): Testing set | | |
| co2[229:456] - without division in lags | | Abs error of prediction |
| Mean of residuals | 0.1708991 | 0.5816793 |

| co2 (1:1) | x<-co2[1:216] | co2[217:432] Lag=24 x 9 |
|---------------------------------|-------------------------|--------------------------------|
| Training/Testing set | Lag=24 x 9 | |
| Forecast arima(1,1,1) | Mae | Mae |
| Abs error of prediction | 0.3452741 | 0.2722938 |
| Mean of residuals | 0.04391378 | 0.04335611 |

| Auto.arima: Training set | | |
|---|--------------|-------------------------|
| x<-co2[1:216] - without division in lags - Training set | | Abs error of prediction |
| Mean of residuals | -0.005578176 | 0.1938634 |
| Arima (2,1,1): Testing set | | |
| co2[217:432] - without division in lags | | Abs error of prediction |
| Mean of residuals | 0.169955 | 0.248319 |

| Taylor (1:1) | x<-taylor[1:1680] | x<-taylor[1681:3360]Lag= |
|---------------------------------|-----------------------------|------------------------------------|
| Training/Testing set | Lag=5 x 336 | 5 x 336 |
| Forecast arima(2,0,2) | Mae | Mae |
| Abs error of prediction | 188.4699 | 188.4699 |
| Mean of residuals | 3.590077 | 3.129715 |

| Auto.arima: Training set | | |
|---|-----------|-------------------------|
| x<-taylor[1:1680] without division in lags- Training set | | Abs error of prediction |
| Mean of residuals | 0.7800593 | 459.88 |
| Arima (4,0,4): Testing set | | |
| taylor[1681:3360] - without division in lags | | Abs error of prediction |
| Mean of residuals | 0.5106114 | 103.1996 |

| gas (1:1) | x<-gas[1:216] | x<-gas[217:432] Lag=24 |
|---------------------------------|-------------------------|----------------------------------|
| Training/Testing set | Lag=24 x 9 | x 9 |
| Forecast arima(1,1,0) | Mae | Mae |
| Abs error of prediction | 185.4429 | 741.2236 |
| Mean of residuals | 35.64954 | 117.0666 |

| Auto.arima (2,1,2): Training set | | |
|---|----------|-------------------------|
| x<-gas[1:216] - without division in lags - Training set | | Abs error of prediction |
| Mean of residuals | 32.97419 | 1061.73 |
| Arima (2,1,2): Testing set | | |
| gas[217:432] - without division in lags | | Abs error of prediction |
| Mean of residuals | 283.3677 | 1788.014 |

In each of the three group of experiments we divided the data into training set / testing set 1:1. After that, our algorithm find the arima model parameters (p, d, q, P, D, Q, constant, mean (aic)) for the entire time series of data which are identical for each lag, and which at the same time give the minimum mean value of AIC. With these output parameters of our algorithm the absolute predictive and the mean value of the residual is calculated. It was the second part of our algorithm. The results are shown in the table.

In order to be able to compare the results of our experiment, the third part of our algorithm calls *auto.arima* function and through it we get the predictive and mean value of the residual for all sets of data (training set) without dividing in lags.

As can be seen from the table 1. (for a group of data co2 with a lag=12 in the training set) residual values vary from 0.02405235 with our algorithm to -0.005477198 with *auto.arima*, and the predictive value from 0.4373796 with our algorithm to 0.1242049 with *auto.arima*.

In the taylor data group (Training set with a lag=336) the mean residual value changes with our algorithm the absolute error of prediction changed from 188.4699 to 459.88 with the *auto.arima* function, which is significantly larger deviation with the *auto.arima* function.

The biggest differences are apparent in the fourth group of data (gas) where lag=24 and where in the training set the mean residual value with our algorithm changed from 35.64954 to 32.97419 (little smaller with *auto.arima* function), while the absolute prediction error increased from 741.2236 to 1061.73 with the *auto.arima* function.

Another important observation is that the forecasting accuracy of the ARIMA model diminishes gradually at this stage of the growth process, from period to period. Even though the forecasting power of the methodology seems

limited, it should be taken into consideration that the forecasting improvement is for one day or one-month horizon. This single day or month's improved forecast could make the difference in the sense of competition, as this knowledge is a useful guideline for the upcoming day or month's strategy programming.

5. Conclusions

This paper presented a new methodology that delivers short-term forecasts of the natural phonemes measured for a period of time in the spatial distributed sensors that measure different parameters. After obtaining enough actual data to construct a time-series, a diffusion model and an ARIMA model are applied over the sample and the first forecasts of the ARIMA model are used to perform an improved short-term prediction using a conventional aggregate model. Since the two categories of modelling are of completely different concepts and implementations, the choice of combining their forecasts in the most suitable manner has been made. Therefore, it can be concluded that our algorithm that divides the time series into lags and calculates the best parameter values of the ARIMA model ($p, d, q, P, D, Q, constant, mean(aic)$), valid at the same time for each lag, gives more accurate and significantly better solutions in forecasting procedures than the *auto.arima* function that calculates the best model parameters by treating the whole time series (table 1-8) for several data set.

This methodology can be probably applied over all cases for obtaining future forecasts. Its main limitations consist of the prerequisite for having enough historical data points in order to create a time-series and that the diffusion process should be at the time point when the take off stage of the diffusion process is initiated. The study was limited to a forecasting horizon of one day or one month ahead. Future research in this topic includes the application of the methodology in other cases of space model forecasting, as well as the further investigation of its use in other stages of the process and for other forecast elements.

The accuracy of each approach's forecast is measured in terms of MAE in our research and applications, as noted earlier in this paper. This approach have been also compared based on the other two measures of accuracy, the Mean of residuals Error (MRE) and the Mean Absolute Error (MAE) in order to confirm the superiority of our approach opposite of the *auto.arima* model and this observation of the other measures resulted in the same conclusions as well.

Our intention in future research in this topic will include the application of the methodology in other cases of space model forecasting, as well as further investigation of its use in other stages of the process and for other forecast elements. As a future work, we plan to investigate a combination of the multivariate time series forecasting technique and polynomial regression with a combination of linear interpolation techniques like inverse distance

weighting or Kriging that include not only temporally but also spatially distributed data.

References

- [1] J.S. Armstrong, "Combining forecasts: the end of the beginning or the beginning of the end?", *Int. Journal Forecast.* 5, 1989, pp.585–588,
- [2] G.E.P. Box, G.M. Jenkins, *Time series analysis: Forecasting and control*. Holden Day, San Francisco, 1970.
- [3] P.J. Brockwell, R.A. Davis, *Time Series: Theory and Methods*, 2nd ed. Springer, 2009.
- [4] D.W. Bunn, "Combining forecasts", *European Journal of the Operational Reseach.* 33, 1988, pp. 223-229.
- [5] N.A. Gershenfeld, A.S. Weigand, "The Future of Time Series", *Learning and Understanding. Time Series Prediction. Forecasting the Future and Understanding the Past*. In: Eds. A.S.Wigand and Gersehenfeld, N.A. *SFI Studies in the Sciences Complexity*, vol. IV, Addison Wesley, 1993, pp. 1-70.
- [6] S.L. Ho, "The use of ARIMA models for reliability forecasting and analyses", *Computers and industrial engineering.* 35 (1-2), 1998, pp. 213-216.
- [7] R.J. Hyndman, *Data from the M-Competitions*. R package version 1.11, <http://CRAN.R-project.org/package=forecasting>, 2008.
- [8] R.J. Hyndman, M. Akram, B.C., Archibald, "The Admissible Parameter Space for Exponential Smoothing Models", *Annals of the Institute of Statistical Mathematics*, 60 (2), 2008, pp. 407-426.
- [9] R.J. Hyndman, Y. Khandakar, "Automatic time series forecasting", *The forecast package for R. Journal of Statistical Software*, 26(3), 2008.
- [10] R.J. Hyndman, "Data Sets from Forecasting: Methods and Applications By Makridakis", *Wheelwright & Hyndman 1998*, R package version 1.11.<http://CRAN.R-project.org/package=forecasting>, 2008.
- [11] R.J. Hyndman, "Forecasting Functions for Time Series", R package version 1.11, <http://CRAN.R-project.org/package=forecasting>, 2008.
- [12] P. Newbold, "ARIMA model building and the time-series analysis approach to forecasting", *Journal Fore cast.* 2, 1983, pp. 23–35.
- [13] J.W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing", *Journal of the Operational Research Society.* 54, 2003, pp. 799-805.
- [14] A. Timmermann, *Chapter 4: forecast combinations*. *Handbook. Econ. Forecast.* 1, 2006.
- [15] G.U. Yule, "On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers", *Philos. Trans. Roy. Soc. London Ser. A* 226, 1927, pp. 267–298.
- [16] O. Ohashi, L. Torgo, *Wind speed forecasting using spatio-temporal indicators*. In L. D. Raedt, C. Bessiere, D.

- Dubois, P. Doherty, P. Frasconi, F. Heintz, P. J. F. Lucas, editors, 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012), SystemDemonstrations Track, volume 242 of Frontiers in Artificial Intelligence and Applications, pp. 975–980. IOS Press, 2012.
- [17] W. Tobler., A computer movie simulating urban growth in the Detroit region". *Economic Geography*, 46(2), 2012, pp. 234-240.
- [18] M. F. Goodchild, *Spatial autocorrelation*. Norwich, England: GeoBooks. 1986.
- [19] Y. C. Lee, L. Tong, Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge- Based Systems*, (24), 2011, pp. 66–72.
- [20] P.A.P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika*, 37, 1950, pp. 17–23.
- [21] C.D. Keeling, T. P. Whorf, "Scripps Institution of Oceanography (SIO)", University of California, La Jolla, California USA 92093-0220.