# Forecasting Algorithm Adaptive Automatically to Time Series Length

Kolyo Onkov, Georgios Tegos



# Forecasting algorithm adaptive automatically to time series length

Kolyo Onkov[1], Georgios Tegos[2]

[1]Department of Mathematics, Computer Science and Physics, Agricultural University, 12 Mendeleev, 4000 Plovdiv, Bulgaria,
kolonk@au-plovdiv.bg

[2]Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, P.O. Box 14561, 54101, Greece,
gtegos@gen.teithe.gr

**Abstract.** The developed forecasting algorithm creates trend models based on varying length time series by eliminating its oldest member. The constructed criterion evaluates the attained models through estimating the ratio between the average of the stochastic errors for the forecasted period and the average of real values. The best model and forecasting are automatically achieved in contrast to statistical software systems SPSS, STATISTICA, etc. where this process is accomplished progressively by the user. Therefore, this forecasting algorithm is adaptive to the length of time series. Component oriented approach has been used for software implementation. Simulation experiments have been carried out to test the forecasting algorithm using the multidimensional time series database on fishery in Greece. This algorithm is more efficient in case forecasting is applied on large number of time series because it saves time and efforts.

**Keywords:** varying length time series, automatic model fitting, criterion, adaptive algorithm

## 1 INTRODUCTION

Time series forecasting refers to the process of identifying past relationships and trends in historical data for predicting future values [1]. Forecasting is important in time series analysis because it plays a central role in management since it precedes decision making [2]. There are many time series analysis techniques related to forecasting [3]. Trend modelling can be used for forecasting if it is assumed that the studied event will follow the same rules during the historical and the forecasted period. The attained forecasts are a simple consequence of the trend extrapolation. The advantages of forecasts based on trend modelling compared to other forecasting methods are: a) the preliminary evaluation of the forecast stochastic error by trend models and b) the determination of the confidence intervals. As a result, the critical limits, on which the real values of the studied event fluctuate during the forecasted period, can be evaluated.

While modeling the trend of time series for forecasting purposes, the issue concerning the length of the period, within which the trends will be studied, frequently arises. Here there are two tendencies opposing each other. The stochastic error of the forecast decreases in longer historical periods. However, by covering a longer historical period, there is a risk for the model to include the influence of factors that used to function in the past, but later disappeared.

Forecasting time series based on trend modeling is often implemented through using statistical packages such as SPSS, Statistica etc. However, for reapplying trend modeling to different length time series and achieving new forecasted values the same procedure is followed manually. In general, in case of forecasting based on large numbers of time series, automatic forecasting is more efficient [2], [4]. Two automatic forecasting time series algorithms based on state space and ARIMA models have been implemented as R packages for statistical computing [5]. Fuzzy time series models and automatic forecasting techniques are developed to improve forecasting accuracy [6], [7].

This work aims to present an adaptive to time series length algorithm in order to achieve the best models and forecasted values automatically. The idea is to reapply forecasting by decreasing time series length and testing different trend models. As a result, different forecasts are achieved giving the opportunity to choose the most proper one based on objective criterions.

## 2 MATERIALS AND METHODS

The algorithm is based on the gradual elimination of data from the start of the historical time series by applying the same trend model to each resultant time series. The confidence intervals are evaluated according to the different types of trend models on the grounds of which the forecast is achieved.

Let's denote:

$y_i$ (i=1, 2,…, *n*) – time series with length *n*,

$\hat{y}_{n+j}$ – forecasted values, j = 1, 2, ……, L, L – length of the forecasted period

$$S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{– standard error,} \qquad \text{I}$$

$\hat{y}_i$ *(i=1,2,..., n)* – smoothed values by the model

The linear $y = A_0 + A_1 t$ and the polynomial second degree $y = A_0 + A_1 t + A_2 t^2$ trend models are applied to varying length time series. The stochastic error of the forecasted value $\hat{y}_{n+j}$ for linear (equation II) and for polynomial second degree (equation III) model is [8]:

$$e_{\hat{y}_{n+j}} = \frac{S_y}{\sqrt{n-1}} \cdot \sqrt{1 + \frac{1}{n} + \frac{3(n+2j-1)^2}{n(n^2-1)}} \qquad \text{II}$$

$$e_{\hat{y}_{n+j}} = \frac{S_y}{\sqrt{n-1}} \cdot \sqrt{1 + \frac{1}{\sum_{p=1}^{j} t_p^2} \cdot t_{n+j}^2 + \frac{\sum_{p=1}^{j} t_p^4 - (2\sum_{p=1}^{j} t_p^2) t_{n+j}^2 + n \cdot t_{n+j}^4}{n\sum_{p=1}^{j} t_p^4 - (\sum_{p=1}^{j} t_p^2)^2}} \qquad \text{III}$$

where $S_y$ is the standard error by equation I.

The confidence interval for the linear and polynomial second degree models is evaluated as follows:

$$\hat{y}_{n+j} - t_{[1-\alpha]} e_{\hat{y}_{n+j}} \le \tilde{y}_{n+j} \le \hat{y}_{n+j} + t_{[1-\alpha]} e_{\hat{y}_{n+j}} \qquad \text{IV}$$

$\tilde{y}_{n+j}$ – the attained value of y for the j[th] year of the forecasted period;

$t_{[1-\alpha]}$ – Student's test at a level of significance α and n-2 degrees of freedom.

From equations II and III it is noticed that the value of the stochastic error depends directly on the length of time series n to which the trend model is applied and on the length of the forecasted period j. When n increases the value of stochastic error decreases while when j increases the stochastic error also does. Stated in other words, the greater the historical period is, the smaller the stochastic error becomes and on the other hand, the longer the forecasted period is, the greater the stochastic error becomes.

A criterion is created in order to evaluate the forecasts attained from time series with different length. The criterion is related to the stochastic error (equation II and III). During the elimination of data of time series by decreasing each time the value of n by 1, the ratio between the average of the stochastic errors for the forecasted period and the average of real values are computed according to the next equation:

$$C_1(d) = \left( \frac{1}{L} \sum_{j=1}^{L} e_{\hat{y}_{d+j}} \right) \Big/ \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right), \quad \text{d=n, n-1,…,}m \qquad \text{V}$$

*m* is the minimal value of time series length, ($m < n$). The best trend model and forecasting are achieved for the minimal value of $C_1(d)$ criterion when varying the length time series.

The created forecasting algorithm, based on trend models applied to varying length time series, is presented here for one time series by the following steps:

1. Implement steps 2-6 for both trend models;
2. Procedure for trend modeling;
3. Procedure for trend model adequacy by F-test. If the model is adequate then go to step 4 else go to step 6;
4. Computation of smoothed values $\hat{y}_i$, i=1, 2, …$n$ and Standard error $S_y$;
5. Procedure for estimation and storing of:

- Forecasted values $\hat{y}_{n+1}$, $\hat{y}_{n+2}$, …, $\hat{y}_{n+L}$;
- Stochastic errors $e_{\hat{y}_{n+1}}$, $e_{\hat{y}_{n+2}}$ …, $e_{\hat{y}_{n+L}}$;
- Confidence intervals by equation IV applying Student's test;
- $C_1(n)$ value.

6. Next operation;
7. $n=n-1$;
8. If $n < m$ then go to step 10 else go to step 9;
9. Elimination of the most distant year from time series and go to step 1;
10. Application of the criterion for the best forecasting – MIN ($C_1(n)$, $C_1(n-1)$, …, $C_1(m)$) for both trend models;
11. Evaluation and printing the best forecasting results based on both trend models.

The algorithm is adaptive to time series length and as a result it gives the ability to achieve automatically forecasts based on trend modelling. The perspective of this algorithm is to apply different weights on the time series members used for trend modeling. The study of this approach will reveal abilities for attaining more precise forecasting values.

# 3 RESULTS AND DISCUSSION

## 3.1 Software and simulation experiment

Component oriented approach and VB programming language have been used for software implementation according to the algorithm. The procedures are placed in two modules:

- The first of them is consisted of procedures for estimating the descriptive statistics (mean value, standard deviation, standard error etc) as well as for evaluating the F-test and Student test;
- The second one contains the procedures for trend modelling, managing the length of time series, stochastic error, confidence intervals, application of the criterion and presentation of the results.

Generally, the procedures in the second module, "call" the procedures from the first one. VB programming code is also created for managing Excel-sheets which are proper for storing tables containing automatically accessed theoretical values of F-test

and Student test by the trend modelling and forecasting and forecasting results as well.

A simulation experiment has been carried out to test the forecasting software using multidimensional time series database on sea fishery in Greece. Data in database concerns quantities and values of fish catch by areas, fish species, fishing tools and category as well as months, kinds of fishery and employment for the period 1990-2011 [9], [10]. Each time series extracted from FTS database consists of 22 members. For the purpose of forecasting it is divided into two parts: the first twenty members are used as historical time series and the last two (2010, 2011) are compared with the obtained forecasted values for the same years (proportion 20:2).

### 3.2 Forecasting results

The forecasting results presented here are oriented to time series on catch quantity of 71 sea fish species statistically registered in Greece. Table1 presents the part of the obtained results concerning only 10 fish species. Generally, for forecasting purposes the optimal time series length on the particular fish species is different. Besides the type of trend model used for the forecasts is also different. From 71 fish species there are 6 without adequate trend because of big random factor and as a result no forecasts are proposed by the algorithm. Almost all these cases concern fish species with small amount of catches – mean value for the studied period of time do not exceed 200 metric tons (Sprat, Skipjack etc).

**Table1.**Trend model and optimal length time series

| Fish code | Fish name | Trend model | Optimal length time series |
|---|---|---|---|
| 15 | Bogue | Linear | 12 |
| 23 | Goatfish | Polynomial 2 degree | 17 |
| 25 | Red bream | Polynomial 2 degree | 10 |
| 31 | Red mullet | Linear | 12 |
| 35 | Bonito | Polynomial 2 degree | 15 |
| 36 | Sprat | No adequate model | - |
| 41 | Skipjack | No adequate model | - |
| 43 | Goldline | Linear | 13 |
| 55 | Tune fish | Linear | 16 |
| 68 | Cuttle fish | Polynomial 2 degree | 18 |

Table 2 presents more detailed results of the forecasting algorithm applied on quantity of catch for fish species “Bogue”, based on linear trend modeling. For each varying length historical time series (n = 15, 14, 13, 12, 11) the forecasted values for the years 2010 and 2011 as well as the confidence interval (equation IV) and the C1 criterion (equation V) are also presented. The relative error shows the percentage

deviation between real and forecasted values. The best forecast is obtained for time series length 12 when C1 criterion has the minimum value (Figure1). The obtained relative error values are considered satisfactory – 10.87% and 2.47% for the years 2010 and 2011 respectively.

**Table2.** Forecasting results for fish species Bogue by automatic application of linear trend model to varying length time series

| Time series length | Year | Fore-casted values | Confidence interval | | Real values | $C_1$ | Relative error % |
|---|---|---|---|---|---|---|---|
| 15 | 2010 | 2882.92 | 2464.92 | 3300.92 | 3201.69 | 14.50 | 9.96 |
| | 2011 | 2683.34 | 2255.65 | 3111.03 | 3405.40 | 15.94 | 21.20 |
| 14 | 2010 | 3052.99 | 2645.02 | 3460.96 | 3201.69 | 13.36 | 4.64 |
| | 2011 | 2885.30 | 2466.59 | 3304.00 | 3405.40 | 14.51 | 15.27 |
| 13 | 2010 | 3311.15 | 2993.61 | 3628.69 | 3201.69 | 9.59 | 3.42 |
| | 2011 | 3095.09 | 2867.97 | 3522.21 | 3405.40 | 10.24 | 9.12 |
| **12** | **2010** | **3549.77** | **3363.76** | **3735.78** | **3201.69** | **5.24** | **10.87** |
| | **2011** | **3484.84** | **3292.34** | **3677.34** | **3405.40** | **5.52** | **2.33** |
| 11 | 2010 | 3579.59 | 3373.36 | 3785.82 | 3201.69 | 5.76 | 11.80 |
| | 2011 | 3521.54 | 3306.89 | 3736.19 | 3405.40 | 6.10 | 3.41 |

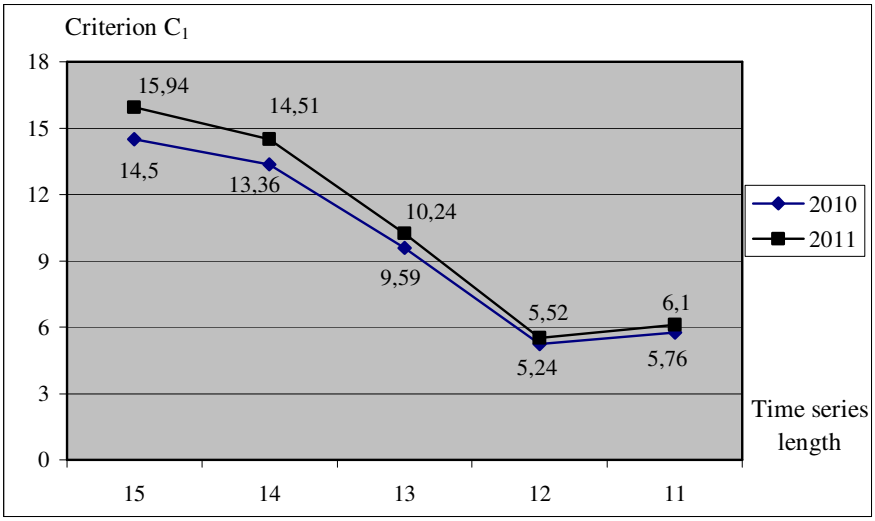


**Figure1.** Values of $C_1$ criterion between 15-11 length time series for fish species "Bogue" in Greek fishery

Table 3 presents the results of the forecasting subroutine application, based on polynomial second degree trend model, on the "total fish quantity" time series. It is evident that the best forecasting result is achieved for time series length 9, because the value of criterion $C_1$ is the smallest – 2.96 and 3.75, for years 2010 and 2011, respectively.

**Table3.** Forecasting results for total quantity of fish catches in Greece

| Time series length | Year | Forecasted values | Confidence interval | | Real values | $C_1$ | Relative error % |
|---|---|---|---|---|---|---|---|
| 20 | 2010 | 61479.50 | 50465.45 | 72493.55 | 70122.20 | 17.91 | 12.33 |
| | 2011 | 55345.39 | 43955.74 | 66735.04 | 62871.50 | 20.58 | 11.97 |
| 19 | 2010 | 69345.30 | 59130.43 | 79560.17 | 70122.20 | 14.73 | 1.11 |
| | 2011 | 66207.68 | 55611.86 | 76803.50 | 62871.50 | 16.00 | 5.31 |
| 18 | 2010 | 78486.68 | 69918.60 | 87054.77 | 70122.20 | 10.92 | 11.93 |
| | 2011 | 79005.61 | 70086.60 | 87924.63 | 62871.50 | 11.29 | 25.66 |
| 17 | 2010 | 86653.35 | 79761.13 | 93545.58 | 70122.20 | 7.95 | 23.57 |
| | 2011 | 90610.88 | 83407.19 | 97814.57 | 62871.50 | 7.95 | 44.12 |
| 16 | 2010 | 92217.97 | 86049.52 | 98386.42 | 70122.20 | 6.69 | 31.51 |
| | 2011 | 98648.66 | 92171.04 | 105126.28 | 62871.50 | 6.57 | 56.91 |
| 15 | 2010 | 89754.04 | 83290.40 | 96217.68 | 70122.20 | 7.20 | 28.00 |
| | 2011 | 95025.24 | 88200.23 | 101850.25 | 62871.50 | 7.18 | 51.14 |
| 14 | 2010 | 91224.82 | 84201.31 | 98248.32 | 70122.20 | 7.70 | 30.09 |
| | 2011 | 97231.40 | 89767.06 | 104695.74 | 62871.50 | 7.68 | 54.65 |
| 13 | 2010 | 89050.90 | 81438.80 | 96663.00 | 70122.20 | 8.55 | 26.99 |
| | 2011 | 93898.06 | 85745.68 | 102050.44 | 62871.50 | 8.68 | 49.35 |
| 12 | 2010 | 80442.93 | 75340.46 | 85545.40 | 70122.20 | 6.34 | 14.72 |
| | 2011 | 80371.25 | 74855.87 | 85886.62 | 62871.50 | 6.86 | 27.83 |
| 11 | 2010 | 79141.56 | 73427.51 | 84855.61 | 70122.20 | 7.22 | 12.86 |
| | 2011 | 78269.03 | 72022.79 | 84515.28 | 62871.50 | 7.98 | 24.49 |
| 10 | 2010 | 71699.59 | 69295.85 | 74103.32 | 70122.20 | 3.35 | 2.25 |
| | 2011 | 65865.75 | 63201.50 | 68529.99 | 62871.50 | 4.04 | 4.76 |
| **9** | **2010** | **69385.92** | **67329.52** | **71442.32** | **70122.20** | **2.96** | **1.05** |
| | **2011** | **61869.42** | **59550.33** | **64188.50** | **62871.50** | **3.75** | **1.59** |
| 8 | 2010 | 68904.13 | 66429.77 | 71378.50 | 70122.20 | 3.59 | 1.74 |
| | 2011 | 61002.20 | 58149.52 | 63854.88 | 62871.50 | 4.68 | 2.97 |
| 7 | 2010 | 66870.92 | 64342.34 | 69399.50 | 70122.20 | 3.78 | 4.64 |
| | 2011 | 57161.69 | 54161.67 | 60161.71 | 62871.50 | 5.25 | 9.08 |

Figure2 presents the real values for total fish catch quantity for time period 1999-2011 and the forecasted values based on time series length n=11, 9 and 7 (polynomial second degree trend model). The results have the following characteristics:

- Forecast of length 11 is optimistic because trend model uses much bigger values of catch quantity at the beginning of the period 1999–2009 than at the end. For the opposite reason forecast of length 7 spanning the period 2003–2009 is pessimistic.
- The most important result consists of the fact that the algorithm automatically finds the optimal time series length 9 and the polynomial second degree model for achieving the best forecast.

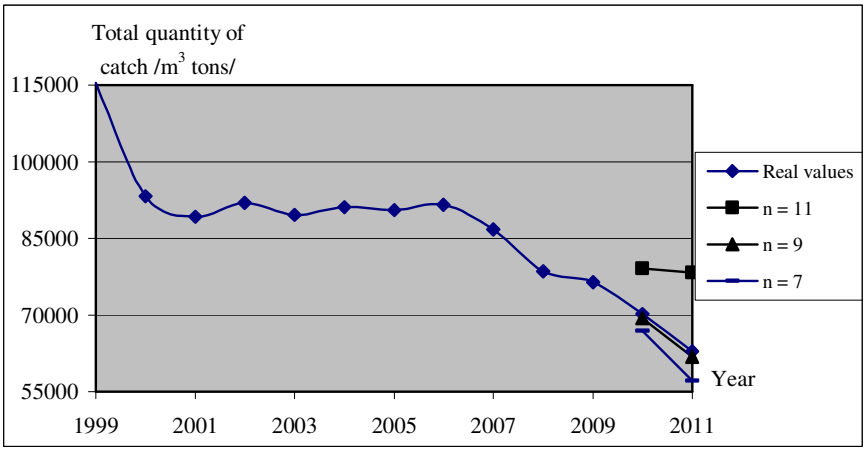


**Figure2.** Forecasting results for varying time series length *n*

The obtained results show that the algorithm is adaptive to the dynamic of the fish catch quantity process and automatically achieves forecasting results.

## 4 CONCLUSION

This work presents an algorithm applied on time series, capable to achieving optimum trend models and forecasting with minimum human intervention. The most significant characteristic of the developed algorithm is the varying length time series used for automatic model fitting. A criterion related to the stochastic error of the forecasted values is constructed for the estimation of the best forecast. When analyzing the set of time series the algorithm behaviour is adaptive to time series length and trend model. As a result, the forecasts are achieved automatically.

A component oriented approach has been used for the implementation of the software. The forecasting algorithm has been tested on time series sets concerning sea fishery in Greece and acceptable forecasting results are attained. The same algorithm can be applied in appropriate time series of other branches giving equivalent results as well. More efficient results are achieved in case forecasting is applied on large numbers of time series because it saves time and efforts.

## REFERENCES


1. Singh, S.: Pattern Modelling in Time-Series Forecasting, An International Journal of Cybernetics and Systems, 31 (1), pp. 49-66 (2000)
2. Makridakis, S., Wheelwright, S., McGee, V.: "Forecasting: Methods and Applications", pp. 926. Wiley, New York (1983)
3. Leonard, M.: "Large-Scale Automatic Forecasting: Millions of Forecasts", SAS Institute, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.9518&rep=rep1&type=pdf
4. Marahaj, E. A., Inder, B.: Forecasting Time Series from Clusters, Working Paper 9/99, Monash University, Australia, http://ideas.repec.org/p/msh/ebswps/1999-9.html
5. Hyndman, R.J., Khandakar, Y.: Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, 27 (3), pp. 1-22 (2008)
6. Wong, W.K., Bai, E., Chu, A.W.: Adaptive Time-Variant Models for Fuzzy-Time-Series Forecasting, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 40 (6), pp. 1531 – 1542 (2010)
7. Chen, S. M., Tanuwijaya, K.: Multivariate fuzzy forecasting based on fuzzy time series and automatic clustering techniques, Journal of Expert Systems with Applications, 38 (8), pp. 10594–10605 (2011)
8. Velickova, N.: Statistical methods for study and forecasting the development of socio-economic processes, Publishing house "Science & Art", Sofia (1981)
9. Tegos, G.: PC-Information System concerning sea fishery time series in Greece, PhD thesis, Computer science department, Agricultural University, Plovdiv (2005)
10. Onkov, K., Tegos, G.: Management solutions for fish resources in Greece based on analysis of multi-dimensional database, Proceedings of the International Conference "Sustainable landscape planning and safe environment", pp. 95-103 (2012)