RESEARCH ARTICLE

AESTIMATIO, THE IEB INTERNATIONAL JOURNAL OF FINANCE, 2015. 11: 78-91 DOI:10.5605/IEB.11.3

Fuzzy hybrid system for forecasting financial time series

Mena, Hermann
Fuenmayor Viteri, Patricio

▸ RECEIVED: 7 OCTOBER 2014
▸ ACCEPTED: 12 NOVEMBER 2014

Abstract

We propose a fuzzy hybrid system for forecasting time series, based on the automatic fitting method `auto.arima` included in the `forecast` package for R. First, we generate predictions and apply fuzzy clustering to identify patterns and tendencies. Then, using inference criteria on the centers of the clusters we end up with a *mean* forecast. The system allows the inclusion of expert criteria, i.e., the user can set up restrictions on the clustering based on *a priori* knowledge of the time series. This approach can be applied to any financial time series meeting the requirements of Seasonal Autoregressive Integrated Moving Average (SARIMA) models. The proposed method is implemented in R. Numerical tests on series of loans, accounts, and saving accounts demonstrate the efficacy of the method.

Keywords:
Financial Time series, Automatic forecasting, Hybrid systems, Fuzzy clustering, SARIMA models.

JEL classification:
G21, G17, G39.

Mena, H. ✉ University of Innsbruck, Department of Mathematics, A-6020 Innsbruck, Austria. Email: hermann.mena@uibk.ac.at.
Fuenmayor Viteri, P. Banco Pichincha, Quito, Ecuador. Email: patricio.fuenmayor@gmail.com

AESTIMATIO, THE IEB INTERNATIONAL JOURNAL OF FINANCE, 2015. 11: 40-53

Un sistema híbrido difuso **para la predicción de series temporales financieras**

Mena, Hermann
Fuenmayor Viteri, Patricio

Resumen

En este artículo se propone un sistema híbrido difuso para la predicción de series temporales. Dicho sistema está basado en el método de ajuste automático `auto.arima` del paquete `forecast` para R. En un primer momento se generan las predicciones y se identifican patrones y tendencias utilizando técnicas de agrupamiento difuso. Posteriormente, utilizando criterios inferenciales sobre los centros de los conglomerados, se finaliza con una predicción en términos de *media*. El sistema propuesto permite la inclusión de criterios expertos, es decir, el usuario puede establecer restricciones en los conglomerados basadas en el conocimiento a priori de la serie temporal objeto de análisis. El procedimiento puede ser aplicado a cualquier serie financiera que cumpla los requisitos de los modelos estacionales autorregresivos integrados de media móvil. El método propuesto se implementa en R. Se han llevado a cabo contrastes numéricos sobre préstamos, cuentas corrientes y cuentas de ahorro que muestran el buen funcionamiento del método propuesto.

Palabras clave:

Series temporales financieras, predicción automática, sistemas híbridos, *clustering* difuso, modelos SARIMA.

1. Introduction

Modeling and forecasting large numbers of time series is a common task in business and finance. In recent years, different methods using high performance computing techniques have been proposed (see for example, Kovalerchuk and Vityaev, 2002; Halgamuǧe and Wang, 2005; Cížek *et al.*, 2005). The choice of the best model for a given time series is usually based on the optimization criterion chosen as reference information. In practice, the expert criteria do not always agree with the estimated values thus necessitating the search for a model that combines expert criteria as well as an acceptable statistical fit. The computational breakthrough of the last decade allows this task to be handled efficiently. A significant amount of research has been conducted on the search for the optimal model by automated process. In particular, for time series meeting the requirements of Seasonal Autoregressive Integrated Moving Average (SARIMA) models, the *forecast* package for the R system for statistical computation (R Development Core Team, 2008), developed by Hyndman and Khandakar (2008), allows us to model and forecast a given time series. This software package is the basis for the hybrid system proposed in this paper.

We use the term 'hybrid' to describe our system because it combines fuzzy clustering and automatic forecasting. Thus, first we generate predictions and apply fuzzy clustering to identify patterns and tendencies. Then, using inference criteria on the centers of the clusters we end up with a *mean* forecast. The system allows the inclusion of expert criteria, i.e., the user can set up restrictions on the clustering based on *a priori* knowledge of the time series. This use of expert knowledge plays a key role in modeling financial time series.

The paper is organized as follows: in Section 2 we briefly describe SARIMA models and introduce the notation. Then, in Section 3 we discuss automatic forecasting and its implementation in the `forecast` package. In Section 4 we review the fuzzy c-means clustering algorithm and the inclusion of expert criteria. In Section 5, we present the fuzzy hybrid system which is the main contribution of this work. Section 6 contains numerical tests with financial series of loans, accounts and savings accounts. Finally, some conclusions and outlook are presented in Section 7.

2. SARIMA models

A SARIMA model is a probabilistic time series model. It represents an adaptation of an Autoregressive Integrated Moving Average (ARIMA) model to specifically fit seasonal non-stationary time series. Clearly, the construction of a SARIMA model de-

pends on the underlying seasonal nature of the series to be modeled. In modeling, it is important to recognize the presence of seasonal components and to remove them from the model so as not to confuse them with long-term trends.

The classical time series model is given in the decomposition form $X_t = m_t + s_t + \varepsilon_t$, where m_t represents the trend component, s_t the seasonal component and ε_t the random noise, for all t. It is based on the assumption that the seasonal component s_t repeats itself in exactly the same way cycle after cycle. In modeling real data it might be reasonable to assume certain randomness of the seasonal component. The SARIMA models, therefore, can prove particularly useful as they allow for randomness in the seasonal pattern from one cycle to the next.

An appropriate model for a given set of observations $\{X_1, ..., X_n\}$, which exhibit no apparent deviations from stationarity and have a rapidly decreasing auto-covariance function, is an ARMA(p,q) model for the mean-corrected data. A time series $\{X_t\}$ is said to follow an autoregressive moving average of orders p and q and is denoted by ARMA(p,q) if for every t:

$$X_t - a_1 X_{t-1} - a_2 X_{t-2} - \ldots - a_p X_{t-p} = \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \ldots + b_q \varepsilon_{t-q}, \quad (1)$$

where $\{\varepsilon_t\}$ represents a white noise process, e.g., ε_t could be normal with zero mean and variance σ^2, i.e. $\varepsilon_t \sim N(0,\sigma^2)$, and $a_1,..., a_p$; $b_1,..., b_p$ are constants such that (1) is both stationary and invertible. This means that polynomials

$$\phi(z) = 1 - a_1 z - a_2 z^2 - \ldots - a_2 z^p$$

and

$$\psi(z) = 1 + b_1 z - b_2 z^2 + \ldots + a_q z^q$$

of orders p and q respectively, have no common factors. The existence and uniqueness of a stationary solution $\{X_t\}$ is equivalent to the condition that $\phi(z) \neq 0$ for all complex z on the unit circle $|z| = 1$. The invertibility condition is fulfilled if $\psi(z) \neq 0$ for all $|z| \leq 1$.

If B denotes the backward shift operator, $B^k X_t = X_{t-k}$ and $B^k Z_t = Z_{t-k}$, $k = 0, 1,...$, then, model (1) can be represented in the form:

$$\phi(B) X_t = \psi(B) \varepsilon_t . \quad (2)$$

Note, if $\psi \equiv 1$ the time series $\{X_t\}$ in (2) represents an autoregressive process of order p, denoted by AR(p), and if $\phi \equiv 1$ then $\{X_t\}$ reduces to a moving-average process of order q, denoted by MA(q).

Let d be an non-negative integer. A non-seasonal ARIMA(p,d,q) process is given, in general, by:

$$\phi(B)(1-B^d)X_t = c + \psi(B)\varepsilon_t \,. \tag{3}$$

where ψ for and ϕ are polynomials of degrees p and q, respectively. Moreover, it is assumed that $\phi(z)\neq 0$ and $\psi(z)\neq 0$ for $|z|\leq 1$. The process is stationary if, and only if, $d = 0$ and then it reduces to an ARMA(p,q) process.

Now suppose that $\{X_t\}$ is observed to be seasonal of period s. The seasonal ARIMA $(p,d,q) \times (P,D,Q)_s$ process with period s, for every t, is given by:

$$\phi(B)\phi(B^s)(1-B^s)^D(1-B)^d X_t = c + \psi(B)\Psi(B^s)\varepsilon_t \,, \tag{4}$$

where $\Phi(z)$ and $\Psi(z)$ are polynomials of degrees P and Q respectively, each containing no roots inside the unit circle. In practice, D rarely exceeds unity and P and Q are in general less than three. For more details, see Box and Jenkins (1976) and Brockwell and Davis (2002).

3. Automatic forecasting

The term automatic forecasting describes a system that generates a time series model, estimates the parameters and computes the forecast. Automatic forecasting algorithms handle large numbers of time series reducing processing time and increasing productivity. The choice of the best model usually relies on statistical criteria, e.g., Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). As stated in Hyndman and Khandakar (2008), most automatic forecasting algorithms are based on ARIMA or exponential smoothing models.

Exponential smoothing methods have been studied since the 1950s. They are classified in terms of the trend component (none, additive, additive damped, multiplicative, multiplicative damped) and the seasonal component (none, additive, multiplicative), giving a total of fifteen methods. Exponential smoothing methods are optimal state space models (see Hyndman *et al*., 2002, and Ord *et al*., 1997). Note that there are similar state space models for the exponential smoothing variations. Moreover, all the exponential smoothing models can be written as state space models (Hyndman *et al*., 2008).

ARIMA models are considered more general than exponential smoothing. However, there is no ARIMA model equivalent to the non-linear exponential smoothing models. On the other hand, for stationary models, ARIMA models are a better choice. In addition, there are a number of ARIMA models that deal with seasonal data. For this

reason, in this paper we use the ARIMA models implemented in the *forecast* package of Hyndman and Khandakar (2008). As stated by users of ARIMA models, for automatic forecasting it is crucial to choose the order of the model appropriately, i.e., to define the values p, q, P, Q, D and d. If d and D are given, the orders p, q, P and Q can be selected via AIC:

$$AIC = -2\log(L)+2(p+q+P+Q+k),$$

where $k=1$ if $c\neq 0$ and $k=0$ otherwise, and L is the maximized likelihood of the model fitted to the difference data $(1-B^s)^D(1-B)^d X_t$. The likelihood of the full model for X_t is not actually defined and so the AIC values for different levels of differencing are not comparable. For seasonal data, Hyndman and Khandakar (2008) use ARIMA$(p,q,d)\times(P,Q,D)$ models where $D=0$ or $D=1$ depending on an extended Canova-Hansen test. In their implementation they allow any value of $s>1$. For $s>12$ the critical value is obtained in a simpler way.

It should be noted that in applications automated forecasting requires the validation of an expert. Thus, a technique that allows the inclusion of expert criteria is needed. This is the main advantage of our method, which is described in Section 5.

4. Fuzzy *C*-Means clustering

Cluster algorithms identify groups of data based on shared similarities. Measurements of such similarities are clearly defined, e.g., for metric spaces the similarity is defined by means of a distance norm among the data vectors. The cluster centers are not usually known beforehand and are determined by the cluster algorithm while partitioning the data. Most cluster algorithms do not rely on assumptions common to statistical classification methods, such as statistical data distribution; they use instead an objective function to measure the desirability of partitions. Nonlinear optimization algorithms are used to find the local optima of the objective function. These algorithms can be classified according to whether the clusters, as subsets of the entire data set, are fuzzy or crisp. Algorithms based on classical set theory, classify objects according to whether or not they belong to a cluster, which is known as *hard* clustering. Fuzzy clustering algorithms allow objects to belong to several clusters simultaneously, with different degrees of membership (between 0 and 1).

The fuzzy c-means (FCM) algorithm is one of the most popular clustering algorithms. The minimization of the c-means functional is solved by the Picard iteration through the first-order conditions for stationary points of the objective function (see Palit and Popovic, 2005). In order to execute the algorithm, the following parameters have to be specified: the number of clusters c; the fuzziness exponent m

(as m approaches one the partition becomes a hard partition, usually selected as 2); and the tolerance and norm-inducing matrix A (a common choice for A is the identity matrix, and then the norm is the standard Euclidean norm). The most important parameter in the FCM algorithm is the number of clusters, c, as the remaining parameters have little influence on the resulting partition. Note that the algorithm will look for c clusters regardless of whether or not they are actually present in the data. Thus, two main approaches for determining the appropriate number of clusters are typically used: validity measures and iterative merging. Validity measures are scalar indices that assess the goodness of the partition. For the FCM algorithm the Xie-Beni index has been found to perform well in practice (see Xie and Beni, 1991). In the iterative cluster merging, one starts with a large number of clusters and by successively merging clusters that are similar with respect to a criterion, the number of clusters is reduced. Of course, *a priori* knowledge of the data can be used to properly determine the number of clusters, as is the case with the applications in our study. For more information we refer the reader to Palit and Popovic (2005), Ross (2004), Höpper *et al*. (2000) and references therein.

In our model we use the implementation of the FCM algorithm which is available in the package `e1071`, function `cmeans` for `R`.

5. Fuzzy hybrid system

The fuzzy hybrid system (FHS) combines fuzzy clustering and automatic forecasting. It essentially consists of four stages, which internally use functions and specific routines for structured data. The FHS can be applied to *large* systems, i.e., hundreds of sets of monthly data from several decades. In order to visualize the results, graphic functions were also developed. We describe below each stage of the FHS.

STAGE 1:

involves the generation of the model parameters to be tested in the search for the best model. The choice of the best model is based on the minimization of the information criterion (IC) used in `auto.arima2`. This is a straightforward modification of the `auto.arima` function of the `forecast` package. The difference is that the parameter lists resulting from the search for the parameters that minimize the IC are exported to estimate the models. In order to cover a wider range of models, three information criteria are used: AIC, BIC and Akaike second order (AICc). There is a parameter criterion in the main function. A key part of this stage is the exclusion of the models that are considered non-eligible by experts, i.e., those series that do not display the expected behavior, such as trend, seasonality, etc. These models are computed iteratively by the system and they

are stored and incorporated into the process as a filter. Consequently, they are not included in the resulting parameter list.

- **STAGE 2:**

  forecasts the generated models using the `forecast` function. The function requires the input of data series and parameters indicating the orders of the model (p,d,q,P,D,Q), as well as information on whether or not the model includes a drift. The function generates the model parameters, estimates, fittings, residuals, etc. The best model in terms of the AIC is also computed. In addition, the management structures for IC values, which are needed in the next stage, are included in the output list.

- **STAGE 3:**

  analyzes and makes inference using the forecast matrix. First we plot the forecasts and identify the pattern; in this regard the FHS can be seen as a pattern recognition technique. We analyze each month in order to find a value which represents the monthly forecast (median forecasting). We interpret the pattern as similarity or proximity of the elements of the set. For this we use the FCM algorithm allowing a flexible level of ownership. This allows us to efficiently deal with the uncertainty. As explained in Section 4, it is very important to properly define the number of cluster centers. Thus, we define three centers, since the forecasts can be classified as accurate estimation, underestimation and overestimation. Finally, we identify the resulting pattern in each case. The median forecast is based on two weighting criteria for each cluster center. The first criterion considers the number of elements in each group, and the second criterion is based on the number of items in each group which come from low IC models.

- **STAGE 4:**

  plots the results where fittings, residuals, forecasts, median forecast and the best SARIMA model are visualized using the resulting lists of each stage.

R-Script. In the script for R the Fuzzy Hybrid System (FHS) is implemented. The script is developed for the 32-bit version of R and the `forecast`, `sqldf`, `e1071`, `RColorBrewer`, `cwhmisc`, and `dummies` packages are required. The following functions were developed:

- `fhs_mdl` generates the list of objects resulting from the automated modeling, matrix of models and the best models based on the information criteria IC.
- `fhs_prn` generates a list of matrices of fittings, residuals and forecasts. It saves the results of the best model based on the AIC.
- `fhs_cdp` generates fuzzy clusters of forecasts.
- `fhs_grf` generates the graphics.

Detailed information about the script and instructions on how to execute it can be found in the script information file. The script is available in Fuenmayor and Mena (2012).

Note that the model might be different depending on the variable, e.g., in the next section we test the method with series of loans, accounts or savings accounts. Thus, in order to estimate the correct order of the SARIMA model, the ACF and PACF functions should be included.

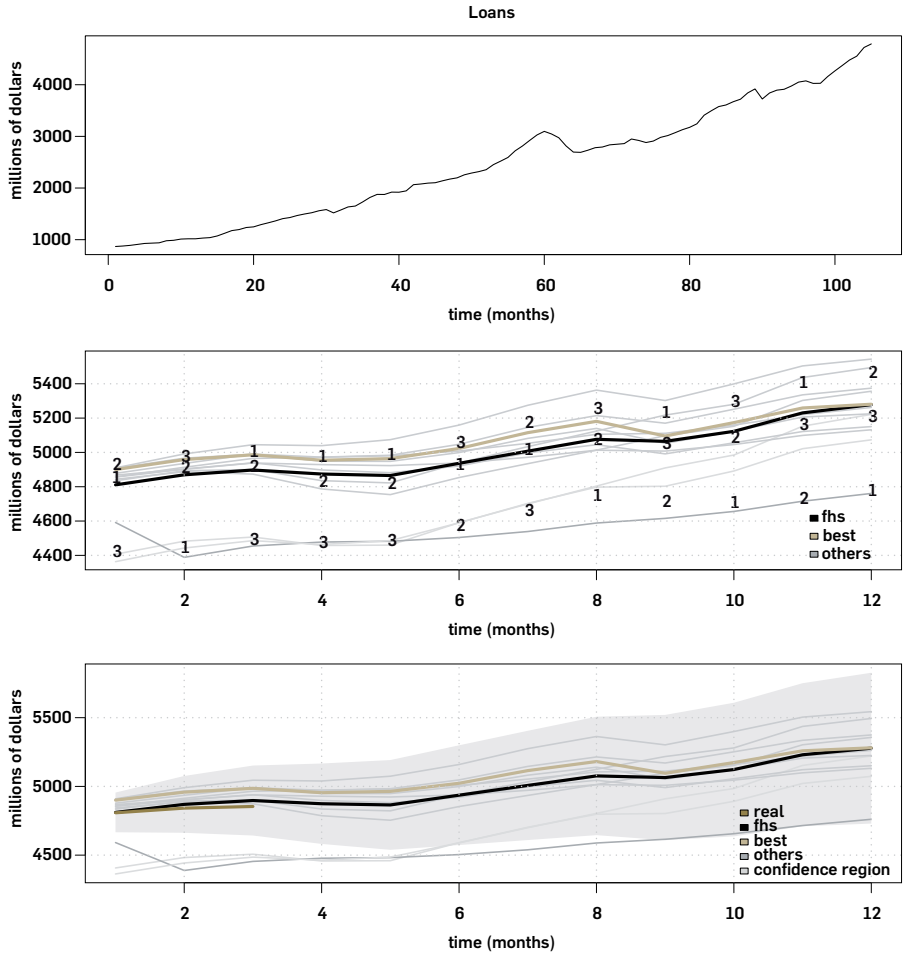
6. Numerical results

The proposed method is tested on three financial series: loans, accounts and savings accounts. We use data from the Ecuadorian bank with the greatest participation in the financial system, provided in the monthly bulletins published by the Ecuador's banking and insurance regulator. The same approach can be used in liquidity scenarios, budgeting, etc. An application on tax forecasting can be seen in Fuenmayor and Mena (2009). These monthly series cover the period from January 2004 to December 2011. The last three months are used to test the forecasts. In order to compare the models, we consider the following goodness-of-fit measures:

- $RMSE = \left(\sum_i^n \frac{e_i^2}{n}\right)^{1/2}$ root mean square error,
- $MAE = \sum_i^n \frac{|e_i|}{n}$ mean absolute error,
- $MAPE = \sum_i^n \left|\frac{e_i}{nx_i}\right|$ mean absolute percentage error.

The evolution of the series for loans is plotted in Figure 1 (top panel). The series show a regularly increasing trend throughout most of the seven years observed. However, they show a decline in the first six months of 2009. At that time there was a great deal of uncertainty sparked by rumors that Ecuador would abandon the use of the dollar as an official currency. Banks consequently stopped giving credit, which explains the decreasing pattern in this period.

Figure 1. **Loans series, inference in forecasts and forecasting tests with real data.**



After the completion of the search process, the model with the lowest AIC can be found:

```
> mdlcar
Series: datcar
ARIMA(2,0,0)(1,1,0)[12] with drift
  Coefficients:
         ar1      ar2     sar1    drift
      1.3270  -0.3647  -0.5656  38.1722
s.e.  0.0959   0.0987   0.0918   8.0665

sigma^2 estimated as 3700:  log likelihood = -517.68, aic = 1045.37
```

The resulting model maintains the increasing pattern and it seems to be linear, see Figure 1 (middle panel). The forecasts together with the results of FHS are also plotted in Figure 1.

Figure 1 (bottom panel) shows the behavior of FHS over the three-month forecast. We defined a confidence zone of $\pm 3\%\sqrt{h}$. The factor $\sqrt{h}$ is considered as uncertainty over time. This factor is also called the rule of the square root of time and it was proposed by RiskMetrics volatilities in forecasting (see Morgan Guaranty Trust Company, 1996). The results show the superior performance of the FHS. Here, as in the (middle panel), *others* means other heuristic models.

Similar tests were performed for Accounts (Figure 2) and Saving Accounts (Figure 3). Table 1 shows the values of the three goodness-of-fit measures for testing the forecast. The FHS, in general, performs better than SARIMA models.

Figure 2. **Accounts series, inference in forecasts and forecasting tests with real data.**

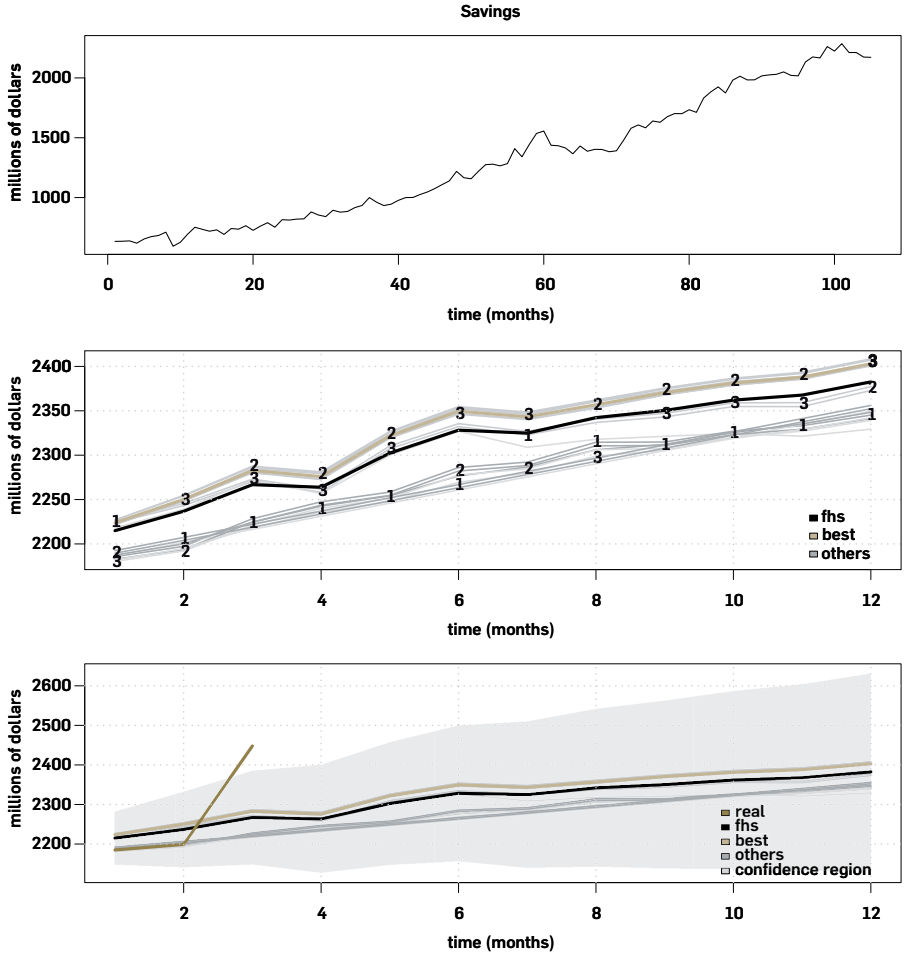


Figure 3. **Saving accounts series, inference in forecasts and forecasting tests with real data.**

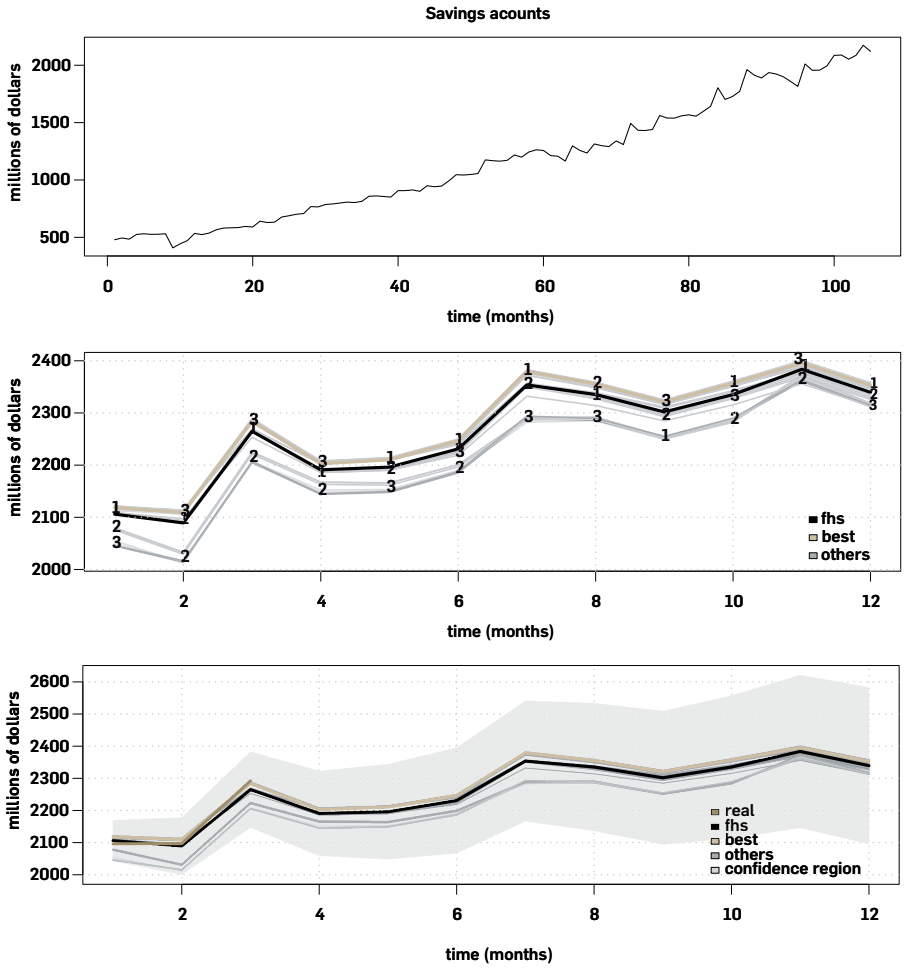


Table 1. **Values of goodness-of-fit measures for testing in the three examples**

	RMSE		MAE		MAPE	
Models	SARIMA	FHS	SARIMA	FHS	SARIMA	FHS
Loans	114.52	33.66	113.33	30	2.34	0.62
Accounts	102.22	108.17	85	83	3.61	3.50
Saving ac.	15.85	18.13	14.67	15.33	0.69	0.69

7. Conclusions and Outlook

In this study, fuzzy clustering and automatic forecast techniques are used to model financial time series meeting the requirements of SARIMA models. The proposed

method can be applied to *large* systems, e.g., hundreds of sets of monthly data from several decades. An additional advantage is that it allows the inclusion of expert criteria in the clustering, which is of great value in business where *a priori* information is usually available. The FHS is a forecasting system; in this sense it is not directly comparable with `auto.arima` or other similar methods. The main idea is to provide experts with an additional forecast. The FHS is implemented in a script for R, which opens up opportunities for further development, e.g., incorporating the modeling of the variance, or new inference criteria, etc. The numerical examples demonstrate that our approach performs well.

Acknowledgement

We would like to thank the referees for their valuable comments. They greatly helped to improve this manuscript.

References

- Box, G. and Jenkins, G. (1976). *Time Series Analysis; Forecasting and Control*, Holden Day, San Francisco.
- Brockwell, P. and Davis, R. (2002). *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, 2nd Edition, Springer Verlag, New York.
- Čížek, P., Härdle, W. and Weron, R. (2005). *Statistical tools for finance and insurance*, Springer-Verlag, Berlin.
- Fuenmayor, P. and Mena, H. (2012). *Fuzzy Hybrid System for Forecasting Time Series- fhs R-script*, http://homepage.uibk.ac.at/~c7021020/, Software section.
- Fuenmayor, P. and Mena, H. (2009). *Inteligencia computacional en la modelación de series financieras: enfoque de la lógica difusa*, Memorias I Congreso Científico Internacional en Economía y Finanzas, EPN, Quito, Ecuador.
- Halgamuge, S.K. and Wang, L. (2005). *Computational Intelligence for Modelling and Prediction*, Springer-Verlag, Netherlands.
- Höpper, F., Klawonn, F., Kruse, R. and Runkler, T. (2000). *Fuzzy cluster analysis*, Wiley, Chichester.
- Hyndman, R.J. and Khandakar, Y. (2008). *Automatic* Time Series Forecasting: The forecast Package for R, *Journal of Statistical Software*, **27**(3), pp. 1-22.
- Hyndman, R.J., Koehler, A.B., Snyder, R.D. and Grose, S. (2002). A State Space Framework for Automatic Forecasting Using Exponential Smooth Methods, *International Journal of Forecasting*, **18**(3), pp. 439-454.
- Hyndman, R.J., Koehler, A.B., Ord, J.K. and Snyder, R.D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*, Springer, Heidelberg. URL http://www.exponentialsmoothing.net

- Kovalerchuk, B. and Vityaev, E. (2002). *Data mining in finance*, Kluwer Academic Publishers, Boston, USA.
- Morgan Guaranty Trust Company (1996). *RiskMetrics Technical Document*, 4th edition, J.P. Morgan/Reuters, New York.
- Ord, J.K., Koehler, A.B. and Snyder, R.D. (1997). Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models, *Journal of the American Statistical Association*, **92**, pp. 1621-1629.
- Palit, A.K. and Popovic, D. (2005). *Computational intelligence in time series forecasting*, Advances in Industrial Control, Springer-Verlag, London.
- R Development Core Team (2008). R: A Language and Enviroment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URLhttp://www.R-project.org/
- Ross T.J. (2004). *Fuzzy logic with engineering applications*, second edition, Wiley, Chichester.
- Xie, X.L. and Beni, G.A. (1991). Validity measure for fuzzy clustering, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **3**(8), pp. 841-846.