ANNALS OF THE UNIVERSITY OF ORADEA
FASCICLE OF TEXTILES, LEATHERWORK

PREDICTING DEMAND FOR COTTON YARNS

SALAS-MOLINA Francisco[1], DÍAZ-GARCÍA Pablo[2]

[1] Hilaturas Ferre, S.A., Les Molines, 2, 03450 Banyeres de Mariola, Spain, E-Mail: francisco.salas@hifesa.com

[2] Universidad Politécnica de Valencia, Ferrándiz y Carbonell, s/n, 03801 Alcoy, Spain, E-Mail: pdiazga@txp.upv.es

Corresponding author: DÍAZ-GARCÍA, Pablo, E-mail: pdiazga@txp.upv.es

Abstract: *Predicting demand for fashion products is crucial for textile manufacturers. In an attempt to both avoid out-of-stocks and minimize holding costs, different forecasting techniques are used by production managers. Both linear and non-linear time-series analysis techniques are suitable options for forecasting purposes. However, demand for fashion products presents a number of particular characteristics such as short life-cycles, short selling seasons, high impulse purchasing, high volatility, low predictability, tremendous product variety and a high number of stock-keeping-units. In this paper, we focus on predicting demand for cotton yarns using a non-linear forecasting technique that has been fruitfully used in many areas, namely, random forests. To this end, we first identify a number of explanatory variables to be used as a key input to forecasting using random forests. We consider explanatory variables usually labeled either as causal variables, when some correlation is expected between them and the forecasted variable, or as time-series features, when extracted from time-related attributes such as seasonality. Next, we evaluate the predictive power of each variable by means of out-of-sample accuracy measurement. We experiment on a real data set from a textile company in Spain. The numerical results show that simple time-series features present more predictive ability than other more sophisticated explanatory variables.*

Key words: *Time-series, random forests, trend, seasonality, futures contracts.*

1. INTRODUCTION

Eliciting the size of inventory buffers for fashion textile products is by no means straightforward. On the one hand, buffers help reduce the lack of synchronization between customers demand and production. On the other hand, a high amount of resources in terms of space and necessary funds are required to maintain stocks. It is also important to know when products are going to be demanded as a way to plan both purchases and production. As a result, predicting demand for products is a crucial task for production managers.

Time-series analysis is a well-developed research field [1]. A time-series is a sequence of observations taken in time sequentially, e.g., a monthly sequence of the quantity of cotton yarns shipped from a factory. The traditional time-series approach is concerned with the building of statistical linear models and their use for forecasting purposes. However, non-linear time-series analysis [2] provides a much wider class of forecasting models and relaxes the common constraints imposed by linear models such as the Gaussianity of forecasting errors.

Demand for fashion products presents a number of particular characteristics such as short life-cycles, short selling seasons, high impulse purchasing, high volatility, low predictability,

tremendous product variety and a high number of stock-keeping-units [3]. These facts make forecasting demand for fashion products a complex task that is not usually well captured by common linear approximations [4]. As a result, recent proposals in predicting demand for fashion products suggested non-linear forecasting models [4,5]. However, these works were based on predicting demand for fashion products based only on past observations.

In this paper, we follow a more general approach by: i) including a number of additional explanatory variables in the forecasting model; and ii) assessing the importance of each explanatory variable. To this end, we rely on random forests [6] in an attempt to capture possible non-linear patterns in forecasting demand for cotton yarns. A random forest is a forecasting technique based on an ensemble of slightly different decision trees. A decision tree is a non-linear model that splits the input space into subsets based on the value of a particular feature. This technique allows us to consider as many explanatory variables as desired such as currencies, cotton prices and any other time-series feature. In addition, random forests provide an assessment of the importance of the input variables used to estimate the model. From this assessment, we are able to identify those variables with more predictive ability and those with no predictive power.

Summarizing, we propose a method to analyze the importance of alternative explanatory variables in predicting demand for cotton yarns based on random forests. We first identify a number of tentative explanatory variables. Next, we use these variables to both estimate the random forest model and assess the importance of variables. In this step, we use a real data set from a textile company in Spain containing monthly demand quantities for cotton yarns from January 2012 to March 2017. Finally, we validate the model by computing the predictive accuracy of the model using a test data set not used in the estimation phase.

The results derived from a numerical example using our real world data set show that simple time-series features present more predictive ability than more sophisticated explanatory variables. These results imply that a number of costless time-series features can be explored as a first step to time-series forecasting.

The structure of this paper is as follows. In Section 2, we describe the methodology used. In Section 3, we present the results derived from a numerical example applying our methodology to a real world data set. Finally, in Section 4, we provide some concluding remarks.

2. METHODOLOGY

In this section, we describe the methodology used to both build and validate a random forest forecasting model for cotton yarns based on a number of explanatory variables.

2.1 Feature extraction

The notion of feature extraction is concerned with the process of identifying and selecting relevant explanatory variables for forecasting purposes. In this paper, an explanatory variable means any data x_i that can be used as an independent variable to predict the value of a dependent variable y by means of a given forecasting function f:

$$y = f(x_1, x_2, ..., x_n) + \varepsilon \quad \textbf{(1)}$$

where n is the number of different explanatory variables and ε is the prediction error. Some particular examples of explanatory variables that can be used for predicting demand for industrial textile products are the exchange rate for currencies involved in commercial transactions or the price of raw materials. We here refer to this class of explanatory variables as *causal variables* due to some expected correlation between the dependent and the independent variable.

An additional kind of explanatory variables available for forecasting when dealing with temporally annotated data is the concept of *time-series feature*, meaning any time-related attribute that can be used for forecasting. Classical time series analysis is based on past observations to produce forecasts (autoregression). Then, we say that past observations are time-series features. The number of time-series features is virtually infinite since we can obtain without much effort a wide variety of statistical attributes from a given time-series. Some time-series features are related to seasonality such as a categorical variable with the month, the week or the day when the observation is taken. Some other time-series features are related to trend attributes such as the rolling-mean of the last three observations. And some other time-series features are related to possible trend such as the variation observed in the last six observations.

Since we experiment on a monthly time-series of demand for cotton yarns, we here rely both on causal variables and time-series features to suggest a set of tentative explanatory variables as described in ***Table 1***.

Table 1: *Tentative explanatory variables*

x_i	Name	Description	Type
x_1	Rate	Average monthly exchange rate EUR/USD	Causal
x_2	Cotton	Average monthly quotation Cotton No. 2 Futures [7]	Causal
x_3	Month	Month of the year of each observations	Time-series
x_4	MA3	Moving average of the last 3 observations	Time-series
x_5	MA6	Moving average of the last 6 observations	Time-series
x_6	Var12	Variation in the last 12 observations	Time-series

2.2 Model selection and out-of-sample validation

We mentioned in the introduction that a random forest is a non-linear forecasting technique based on an ensemble of decision trees. Random forests provide an assessment of the importance of the input variables in terms of mean error decrease provided by each variable. We use this assessment to identify those variables with less predictive ability. To this end, we start with a time-series y with monthly demand for cotton yarns and a data set of explanatory variables x_1 to x_6 as described in ***Table 1***. Following the common out-of-sample model validation procedure [8], we proceed as follows:

1. Separate the entire data set in a training set with the first 80% of the observations and a test set with the remaining 20% for validation purposes. These data sets contain both the time-series y and the explanatory variables x_1 to x_6
2. Estimate the forecasting random forest model using data from the training set.
3. Remove variable x_i with the least predictive ability from the training set.
4. Repeat steps 2 and 3 until only one explanatory variable is left.
5. Select model with the minimum out-of-sample forecasting error e_{out} computed using only data from the test set as follows:

$$e_{out} = \frac{\sum_{\text{test}} (y - f)^2}{\sum_{\text{test}} (y - m)^2} \quad \textbf{(2)}$$

where m is the average of time-series y over the training set. Note that the denominator in equation **(2)** is used as a normalization factor used to discard models with poorer accuracy than a trivial

forecast such as the mean *m* of past observations. Values of e_{out} above one implies that forecast *f* is worse than using the mean as a forecast.

By following the aforementioned steps we are able to select the set of variables to build the forecasting model with best performance in terms of out-of-sample forecasting accuracy. The out-of-sample method measures the generalization power of any forecasting model by computing the accuracy of the model using a data set that was not used to fit the model. This method is used to validate models in similar circumstances to a real deployment in practice. As an additional benchmark, we use the classical Holt-Winters method [9] for seasonal time-series forecasting in its additive formulation.

3. NUMERICAL EXAMPLE

In this section, we apply the methodology described in Section 2 in order to predict demand for cotton yarns using a real world data set from a textile company in Spain. Recall, that the initial data set contains 63 monthly observations for cotton yarns and a number of tentative explanatory variables described in Table 1. Then, following the procedure described in Section 2.2, we divide the whole data set in a training data set with 50 observations and a test set with 13 observations, equivalent to a 80-20% split. We begin estimating a model using all the available explanatory variables over the training set. Then, we progressively remove the variable with less predictive ability according to the importance assessment from the random forest model. At each step, we compute the out-of-sample prediction error by means of equation (2). The results derived from this procedure are summarized in Table 2.

Table 2: *Model selection*

Subset of variables	**Description**	**Out-of-sample error (e_{out})**
$x_1\ x_2\ x_3\ x_4\ x_5\ x_6$	Rate, Cotton, Month, MA3, MA6, Var12	0.205± 0.009
$x_2\ x_3\ x_4\ x_5\ x_6$	Cotton, Month, MA3, MA6, Var12	0.242± 0.020
$x_3\ x_4\ x_5\ x_6$	Month, MA3, MA6, Var12	0.219± 0.014
$x_3\ x_5\ x_6$	Month, MA6, Var12	0.225± 0.017
$x_3\ x_5$	Month, MA6	0.190± 0.012
x_3	Month	0.512± 0.010
-	Additive Holt-Winters method	0.265± 0.000

Due to its particular construction, random forests produce forecasts that are also random. Thus, we perform 100 replicates of the forecasting process to compute its average out-of-sample error and its standard deviation. Note that higher than one out-of-sample errors mean that the particular model performs worse than the mean as a trivial forecast. The results summarized in Table 2 show that even though the causal variables considered in this example do not damage the accuracy of the model, their predictive ability is low. Indeed, both the exchange rate EUR/USD and the cotton quotation are the first removed variables under our proposed procedure for model selection due to their low importance in the model.

As a benchmark, we also present the out-of-sample error achieved by the additive Holt-Winters method for time-series forecasting [9]. In this numerical example, random forests outperform the Holt-Winters method with the exception of the model considering only the month as explanatory variable. Furthermore, the random forest model using the month and the moving average of six months produced the best out-of-sample error results. As a result, we select this model

to predict the future behavior of the demand for cotton yarns. Both the real observations and the forecasts obtained with this model for the test set are depicted in **Fig. 1.** For confidentiality reasons, quantities are normalized by multiplying the original time-series by a correction factor. These results show good generalization power, hence validating the model for forecasting purposes.

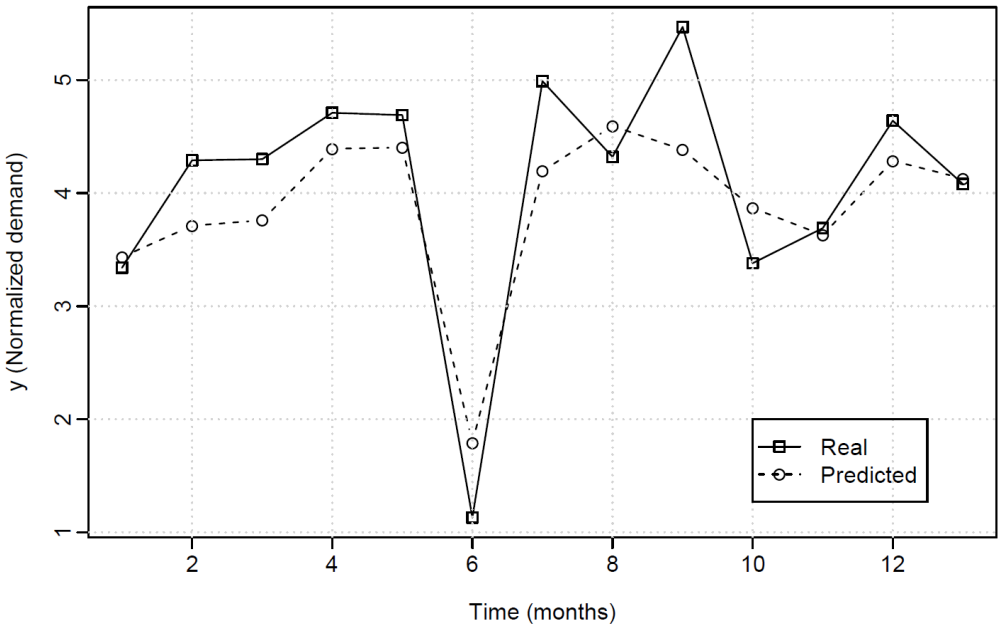

Fig. 1*: Real and predicted demand for cotton yarns over the test set*

4. CONCLUSIONS

In the textile and fashion industry, a flexible forecasting technique is crucial for planning and commercial purposes. In this sense, random forests represent a suitable technique due to both its ability to capture non-linear patterns and the possibility to consider multiple explanatory variables. These explanatory variables are usually labeled either as causal variables, when some correlation is expected, or as time-series features, when extracted from time-related attributes such as seasonality.

In this paper, we rely on random forests to propose a simple methodology to select and validate tentative forecasting models by means of the out-of-sample accuracy measurement procedure. The results derived from a numerical example using our real world data set show that simple time-series features present more predictive ability than more sophisticated explanatory variables. These results imply that a number of costless time-series features can be explored as a first step to time-series forecasting.

REFERENCES

[1] Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). Time series analysis: forecasting and control. 4th Ed. John Wiley & Sons.

[2] Kantz, H., & Schreiber, T. (2004). Non-linear time series analysis (Vol. 7). Cambridge university press.

[3] Nenni, M. E., Giustiniano, L., & Pirolo, L. (2013). Demand forecasting in the fashion industry: a review. International Journal of Engineering Business Management, vol. 5, pp 37-42. Jul. 2013.

[4] Xia, M., Zhang, Y., Weng, L., & Ye, X. (2012). Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. Knowledge-Based Systems, vol. 36, pp. 253-259. Jul. 2012.

[5] Fumi, A., Pepe, A., Scarabotti, L., & Schiraldi, M. M. (2013). Fourier analysis for demand forecasting in a fashion company. International Journal of Engineering Business Management, vol. 5, pp. 30-40, Jul. 2013.

[6] Breiman, L. (2001). Random forests. Machine learning, vol. 45, pp. 5-32. Jan. 2001.

[7] ICE Futures US. Available: https://www.theice.com/products/254/Cotton-No-2-Futures

[8] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting methods and applications. John Wiley & Sons.

[9] Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management science, vol. 6(3), pp. 324-342, Apr. 1960.