A multi-model methodology for forecasting sales and returns of liquefied petroleum gas cylinders

Aldina Correia · Cristina Lopes · Eliana Costa e Silva · Magda Monteiro · Rui Borges Lopes.

Abstract In the liquefied petroleum gas (LPG) cylinder business, one of the most important assets is the LPG cylinder. This work addresses the asset acquisition planning for the LPG cylinder business of a company from the energy sector which has recently started this activity. In order to make the acquisition plan, it was necessary to forecast the sales and the LPG cylinder return rate. For that purpose, an ensemble method using time series techniques, multiple linear regression models and artificial neural networks was employed. Sales forecast were obtained using time series techniques, in particular, moving averages and exponential smoothing. Then, forecast of bottled propane gas sales and return rate was also addressed through multiple linear regression and artificial neural networks. A probability density function was defined for each of the different approaches. Afterwards, using Monte Carlo simulation, the forecast values are obtained by a linear combination of the probability density functions, thus producing the final forecast. Results show that the companys expectation of growth is larger than predicted by the proposed methodology, which means the company should reflect on its current asset acquisition strategy. By combining different approaches, the proposed multi-

A. Correia and E. Costa e Silva
CIICESI – Center for Research and Innovation in Business Sciences and Information Systems and ESTG – School of Management and Technology, Polytechnic of Porto
Felgueiras
E-mail: aic@estg.ipp.pt, eos@estg.ipp.pt

C. Lopes
LEMA – Mathematical Engineering Lab, CEOS.PP – Centre for Organisational and Social Studies of P.Porto, ISCAP – Accounting and Business School, Polytechnic of Porto
Porto, Portugal
E-mail: cristinalopes@iscap.ipp.pt

M. Monteiro
CIDMA – Center for Research & Development in Mathematics and Applications and ESTGA – School of Management and Technology of Águeda, University of Aveiro
Aveiro, Portugal

R. Borges Lopes
CIDMA – Center for Research & Development in Mathematics and Applications and DEGEIT – Department of Economics, Management, Industrial Engineering and Tourism, University of Aveiro
Aveiro, Portugal

model methodology allowed to obtain an accurate forecasting, without requiring a lot of historical data.

Keywords Data analysis · Multivariate analysis · Artificial Neural Networks · Time series analysis · Forecasting · Ensemble method

1 Introduction

In southern European countries, gas is one of the main sources of energy used [32]. Particularly in Portugal, the largest part of gas consumption is for domestic hot water usage, corresponding to 60% of the total gas consumption, from which 65% come from LPG cylinders [14]. LPG cylinders are commonly used for distributing gas when consumption is low or moderate, which is often the case in domestic usage or small businesses (e.g. restaurants). Despite being a very mature or even declining market, it still is a very important business in the Portuguese energy sector.

In Portugal, companies selling LPG cylinders are also responsible for collecting the empty cylinders, regardless of the company from which the previous cylinders were bought (i.e. there is a direct replacement policy). However, as the cylinders are owned by the companies, each competitor can only refill its own cylinders. This makes that companies entering the sector or experiencing growing sales have to purchase additional cylinders to meet demand. These cylinders are expensive, making it the main asset in this business and requiring an adequate planning acquisition strategy if companies are to remain competitive.

This is the problem being addressed in this paper, where a Portuguese company of the energy sector, named ALPHA for confidentiality reasons, wants to find a model to forecast the demand of each type of cylinder. These forecasts are crucial for the company to accurately define an assets acquisition plan, i.e., to determine the amount of LPG cylinders to acquire, and when to acquire them.

To address this challenge, firstly, sales data provided by the company as well as national data were analysed. Then, time series techniques (moving averages and exponential smoothing), multiple linear regression models, and artificial neural networks were used to forecast total propane gas sales and return rates of cylinders (empty bottles). Finally, these methods are combined in a single approach by defining a probability density function for each method and using Monte Carlo simulation to draw values, which are then used in a weighted function (with the weights proportional to the method's accuracy).

Neural networks and their combinations (such as neuro-fuzzy) are widely used in different engineering problem, due to their ability to solve complex problems, as by Naderpour et. al. in [28] and [29].

This work contributes to the literature by combining neural networks with time series techniques and multiple linear regression models in order to increase the accuracy of sales and return rates forecasting. The proposed ensemble method is shown to be most useful in a scenario of rapid growth, where historical data is scarce and often shows high variability. This allows to reduce overfitting and deal with non-linearity and seasonality, thus producing more accurate forecasts than when using a single method [39].

This paper is organized in the following way. The next section is devoted to reviewing the approaches used in literature for these type of problems. Section 3

presents the methodology adopted and some theoretical concepts concerning the forecasting techniques used. Section 4 presents a description of the data and some exploratory analysis. Sales forecast for bottle propane gas for the Portuguese market and for ALPHA company are determined in Section 5 and the forecast of the return rate of the LPG cylinder is addressed in Section 6. Section 7 concludes this paper by summarising the main results and presenting some recommendations.

2 Review of the approaches used in literature

A recent review on forecasting natural gas consumption, mostly applicable in this case, is by Soldo [31] where the author classifies papers in the literature according to forecasting area, forecasting horizon, used data, and forecasting tools. Concerning forecasting tools, the author identifies the following as the most used: statistical models (mostly, time series and regression analysis), artificial neural networks (ANN), and genetic algorithms (GA). Among these tools, at a national/regional level, Soldo suggests the use of classic forecasting tools combined with optimization.

The most common models to forecast sales are time series approaches, either using exponential smoothing methods or autoregressive models. For a more complete introduction on time series see, for example, [5,17]. A moving average is a technique that calculates the overall trend in data and is very useful for forecasting short-term trends. It is the average of a number of time periods and it is called moving because as a new sales number is obtained for a time period, the oldest number in the set falls off, keeping the time period locked. Exponential smoothing focuses more on most recent data, giving more weight to the most recent observations. This type of methods has been extensively used in forecasting [16,19,35, 38], in particular is a widely used method for time series forecast, including sales forecasting [19,24,8]. Simple exponential smoothing, however, does not work well when there is a trend in the data. In these cases, double exponential smoothing can be used to forecast, being the most common example the Holt's method [18]. The method requires separate smoothing constants for the level smoothing factor and for the trend smoothing factor. These time series models forecast future values, taking into account patterns in the historical data and they do not consider any factors that can influence the future. On the opposite, multiple linear regression techniques consider several covariates to model the data, i.e., they can be used to predict future values for the dependent variable given independent variables' data. For a revision in Multiple Linear Regression (MLR) models see [10,15,26,25].

Since 1966 statistical models have been employed for estimation of natural gas consumption [4]. One of the main differences among the works in the literature concerns the forecasting horizon, which often leads to using different approaches. Techniques directed at forecasting annual demand, although being the majority of paper in the literature, are not fitting for this work and therefore are not reviewed here.

Looking at monthly consumption, Liu and Lin [23] use time series (ARIMA) models for forecasting residential natural gas consumption. Addressing data from Taiwan, they identify temperature of service areas and price as the most important influencing factors. Also using ARIMA models, Erdogdu [11] forecasts the future growth in natural gas demand in Turkey, suggesting the model's results are in line

with official projections. Unlike in Liu and Lin's work, Erdogdu concludes that natural gas demand elasticities are quite low, i.e. prices do not influence demand significantly, making consumers more prone to be taken advantage by companies with monopoly power.

Another study focusing on data from Turkey is by Aras and Aras [3] where the authors suggest dividing the year into two seasons, heating and non-heating, and using different models for each season. Three autoregressive time series models are put forward where the deterministic component is a periodic function of time and degree-day (a commonly used heating/cooling measure of how much and for how long the outside air temperature is below/above a certain value).

Focusing on daily demand, Vitullo *et al.* in [36] look at the financial implication of forecasting natural gas, the nature of natural gas forecasting, and the factors impacting its consumption. The authors suggest the most important factors influencing gas consumption are: temperature, prices, wind, demand on the previous month, humidity, precipitation, and luminosity. The authors in [36,35] also present a survey of the mathematical techniques and practices used to model natural gas demand. These authors argue that one of the most common mathematical modelling techniques used to forecast daily demand are MLR. Other works addressing the same time horizon with statistical tools are [37] and [30].

In Vondráček *et al.*, [37], a statistical nonlinear regression model is used, estimating natural gas consumption in individual residences and small businesses based on monthly meter reading data. Sánchez-Úbeda and Berzosa [30], on the other hand, are concerned with daily industrial gas consumption and propose a decomposition model using moving averages, the Linear Hinges Model, and a transitory component to estimate daily variations. Sánchez-Úbeda and Berzosa [30] use degree-day and calendar data (day of the week, holidays, etc.) and are able to predict daily consumption for the following three years.

Finally, in predicting hourly demand for gas, most works employ ANN or GA based approaches. Some of these works, [9,33,34], focus on specific cities from South-eastern European countries and Turkey. Besides historical consumption data, all of them use in their research weather (temperature, wind speed, solar radiation, etc.) and calendar data (day of the week, season, etc.).

Overall, looking at the different works in the literature, gas consumption may be influenced by several aspects, such as, atmospheric temperatures, heliophany (a measure of the day luminosity), wind, relative humidity, rains, minimum and maximum temperatures, demand in previous periods, and prices.

Concerning cylinder returns, Carrasco-Gallego and Ponce-Cueto [7] have addressed returns forecasting techniques, namely dynamic regression models, in the LPG industry. The authors conclude that when a direct replacement policy is in place, the monthly forecast of returns is similar to the monthly forecast of sales, not adding significant value to the use of dedicated forecasting models for returns. Note, however, that in the case addressed by these authors the market share of the company analysed was stable at nearly 80%, thus motivating exploring several different approaches in this work.

Looking at recent forecasting models in the literature, data-driven methods have become more prevalent, some works showcasing different applications are presented hereafter. In Yin *et al.* [40] a fuzzy information granulation approach is proposed and used to forecast freight volume. Feng *et al.* [12] use a multi-model method for wind forecasting, and a review of several data-driven methods

for forecasting energy consumption in buildings can be found in [2]. Furthermore, data-driven models can be used not only in forecasting problems but also in other areas of companies, such as in monitoring and safety control [21].

3 Methodology

The methodology proposed in the present work consists in the combination of several usual techniques used to forecast sales and returns in order to obtain a range of forecast values and its corresponding probability, similarly to Cassettari *et al.* [8]. In their study, as in similar studies, the combination of several forecasting methods has been shown to improve the accuracy and reduce the variability of predictions.

The three forecasting algorithms used are: time series techniques (TS), in subsection 5.1, in particular seasonality coefficients and exponential smoothing; multiple linear regression models (MLR), in subsection 5.2; and artificial neural networks (ANN), in subsection 5.3.

In order to consider the inherent errors of each forecasting technique, each algorithm takes into account that the forecast is described using a normal probability distribution with mean equal to the punctual value of the forecast and variance equal to the mean squared error.

By applying Monte Carlo simulations, in subsection 5.4, a weighted linear combination of the probability density functions (PDF) is used as the output forecast:

$$Y(m) = \alpha_{TS}\, Y_{TS}(m) + \alpha_{MLR}\, Y_{MLR}(m) + \alpha_{ANN}\, Y_{ANN}(m) \tag{1}$$

where $\alpha_{TS}, \alpha_{MLR}, \alpha_{ANN}$ are the weights algorithms, and $Y_{TS}(m)$, $Y_{MLR}(m)$, $Y_{ANN}(m)$ are the forecast values extracted from the PDF of the three forecast algorithms for each month m. It was considered standard weights inversely proportional to the magnitude of the forecast error, in this case MSE (equation (12)). These type of combined methods are usually known as ensemble methods [13]. Fig. 1 resumes the combined methodology. The time series decomposition method is applied, where moving averages are used to produce seasonal indexes. These values are subsequently extracted to estimate and forecast trend through the application of the Holt's algorithm. The use of multiple linear regression models allow to incorporate several variables that can help to explain the behaviour of the variables under study. Thus, in this case, several external variables are used and Figure 1 only presents those that are statistically significant for the study. For the artificial neural network method, a single layer architecture with up to four hidden nodes is used, and the regressors from multiple linear regression are used as predictors of the neural network. Combining these three methods, through the translation of outputs into a PDF, with the definition of weights, which is used to produce a final forecast using Monte Carlo simulation.

One of the advantages of an ensemble methodology is that it produces forecasts that significantly outperform the forecast results of each individual method, since it is able to reduce errors and to improve accuracy between actual values and forecasted values [1,8,39]. Also takes advantage of the characteristics of each method, such as simplicity, interpretation of results and also non-linearity, in the case of ANN method, and reduce the impact of the associated difficulties when

using each isolated method. Time series decomposition methods work well when historical data has a considerable dimension and linear regression method can have problems such as overfitting, strict assumptions and curse of dimensionality while ANN method require a large diversity of training for operation. Thus, combining the three methods allow for the reduction of the impact of the mentioned problems as well as for a better performance, without much computational effort.

In the remaining of this Section some theoretical concepts concerning these forecasting techniques are presented.

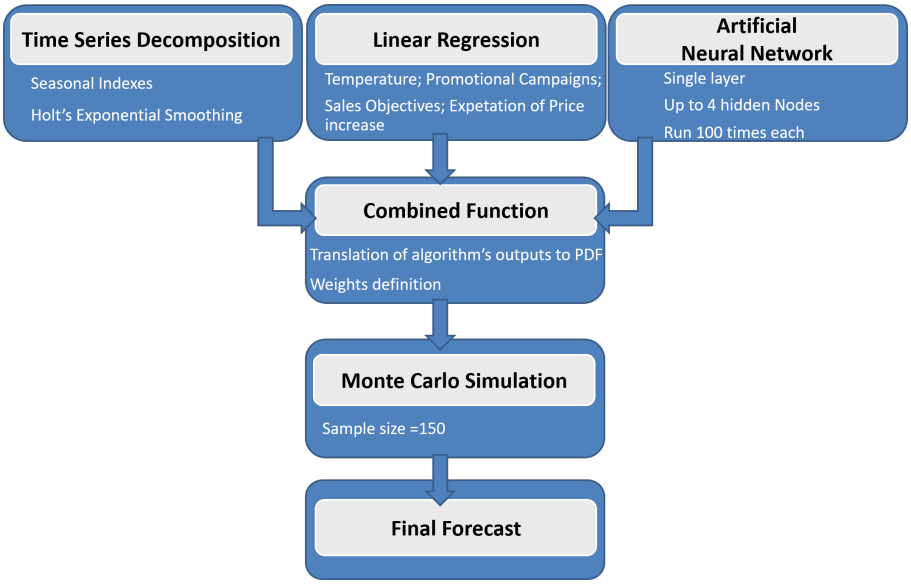

Fig. 1 Overview of the proposed methodology.

3.1 Seasonality Coefficients

To study the seasonality of a time series, seasonality coefficients can be calculated.

The number of observations is given by $N = k \times s$, where k is the number of years and s the number of periods in the year, i.e. months.

Using the centered moving averages method, it is possible to calculate s seasonal indexes, 12 in this particular case (one for each month), that express the amount of sales in each month that are superior (or inferior) to the global mean sales. Using the multiplicative method, the seasonal indexes work as a percentage.

This method starts by, assuming s is an even number, computing the centered moving averages M_t of the series X_t:

$$M_t = \frac{1}{s}\left(\frac{1}{2}X_{t-\frac{s}{2}} + X_{t-\frac{s}{2}+1} + ... + X_{t+\frac{s}{2}-1} + \frac{1}{2}X_{t+\frac{S}{2}}\right), \quad t = \frac{s}{2}+1, \ldots, N-\frac{s}{2}.$$

Then, each value of the series is divided by its centered moving average:

$$S_t^* = \frac{X_t}{M_t}.$$

The non-normalized estimates of the seasonal component at time i of each year:

$$\bar{S}_i = \frac{1}{k-1} \sum_{j=1}^{k} S^*_{i+s(j-1)} \ , \qquad i = 1, 2, \ldots, s. \tag{2}$$

Finally, the standardized estimates of the seasonal components:

$$\hat{S}_i = \bar{S}_i \cdot \frac{1}{\sum_{j=1}^{s} \bar{S}_j} \ , \qquad i = 1, 2, \ldots, s. \tag{3}$$

3.2 Exponential Smoothing

Holt (1957) extended simple exponential smoothing to allow forecasting of data with a trend [20]. This method involves two smoothing equations, one for the level (equation (4)) and one for the trend (equation (5)).

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{4}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} \tag{5}$$

where α, $0 \leq \alpha \leq 1$, and β, $0 \leq \beta \leq 1$, are respectively the smoothing parameters for the level and trend and, at time t, l_t denotes an estimate of the level of the series and b_t denotes an estimate of the trend (slope) of the series.

The forecast combines the last estimated values of the level with the last estimated slope value for the trend:

$$\hat{y}_{t+h|t} = l_t + h b_t. \tag{6}$$

3.3 Multiple Linear Regression Models

Multiple Linear Regression models allow to study the relation between a quantitative response variable (or dependent variable), y, and k regressors (or independent variables), $x_1, x_2, ..., x_k$, so that the model can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon. \tag{7}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients and are constant but unknown, the error ϵ is a random variable with normal distribution. An important objective of regression analysis is to estimate the unknown parameters in the regression model (equation (7)). This process is called fitting the model to the data. This can be done using several parameter estimation techniques, see

Montgomery et. al. in [26] for more detail. The most commonly used is the Ordinary Least Squares (OLS) method, which minimizes the sum of the squares of the random error variables, ϵ_i. The error variables must be mutually independent and normally distributed (i.i.d.), hence they can be seen has a white-noise series (non-correlated with null expected value and constant variance, σ^2). Moreover, the independent variables must be linearly independent from each other (absence of multicollinearity).

After the attainment of the parameters estimates, the following phase of a regression analysis is called model adequacy checking, in which the appropriateness of the model is studied and the quality of the fit ascertained. Through such analysis, the usefulness of the regression model may be determined and the relevant independent variables, that explain the dependent variable behaviour, can be identified. In order to check these adequacy, global ANOVA or F tests and marginal tests are applied, and an analysis of residuals, defined by $e_i = y_i - \hat{y}_i$, should be performed. It is also important to estimate σ^2, the variance of the error term ϵ:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N-k-1} = \frac{SSE}{N-k-1}. \tag{8}$$

The coefficient of determination, or R-Squared (R^2):

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \tag{9}$$

represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The R^2 automatically and spuriously increases when extra explanatory variables are added to the model. Then it is usual to consider an adjusted R^2:

$$R^2_{adj} = 1 - \frac{\sum \varepsilon_i^2 / (N-k-1)}{\sum (y_i - \bar{y})^2 / (N-1)} = 1 - \frac{N-1}{N-k-1}\left(1 - R^2\right) = 1 - \frac{\frac{SSE}{N-k-1}}{\frac{SST}{N-1}} \tag{10}$$

It is a modification due to Henri Theil of R^2 that adjusts for the number of explanatory terms in a model relative to the number of data points. Unlike R^2, the adjusted R^2_{adj} increases only when the increase in R^2 (due to the inclusion of a new explanatory variable) is more than one would expect to see by chance.

Residual analysis implies the verification of several assumptions such as normality, homoscedasticity and the independence of the residuals and the multicollinearity of the independent variables. The normality can be tested using the well known Shapiro-Wilk test, in which the null hypothesis is that the errors have normal distribution. The homoscedasticity is usually tested using the Breusch-Pagan test, whose null hypothesis is the assumption that the variance of errors is constant. The independence of the residuals can be tested using the Durbin Watson test or the Breusch Godfrey test, whose null hypothesis is the mutual independence of the residuals.

In multiple regression models, it is expected to find dependencies between the response variable Y and the regressors. However, if there are also strong dependencies among the independent variables, multicollinearity exists. Multicollinearity can have serious effects on the estimates of the regression coefficients and on the general applicability of the estimated model [27]. To check this it can be considered the variance inflation factor (VIF) for the coefficient β_i:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} = \frac{1}{Tolerance}, j = 1, 2, \cdots, k, \tag{11}$$

where R_j^2 is the coefficient of determination, when the variable X_j is regressed against the other regressors.

Tolerance value is a measure that varies between 0 and 1, being the multicollinearity smaller if the tolerance is closer to 1. Then, when VIF is closer to 0, the multicollinearity is negligible. VIF must be less than 3, and a VIF between 5 and 10 indicates high correlation, that may be problematic. In some situations the limit value 10 is considered.

Several measures of forecast error can be used. All of them share the goal of minimize prediction errors and can be used as criteria for choosing the forecast models. They are also used to estimate the optimal values of the parameters of the models.

The Mean Square Error (MSE):

$$MSE = \frac{1}{N} \sum_{t=1}^{N} e_t^2, \tag{12}$$

then the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{MSE}. \tag{13}$$

The Mean Absolute Error (MAE):

$$MAE = \frac{1}{N} \sum_{t=1}^{N} |e_t|, \tag{14}$$

and the Mean Absolute Percent Error (MAPE):

$$MAPE = \frac{1}{N} \sum_{t=1}^{N} \frac{|e_t|}{|X_t|}. \tag{15}$$

All these measures should be as small as possible.

Several regression models can be computed with a set of independent variables to describe the dependent variable. Thus it is usual to choose the appropriate model using criteria such as AIC, AICc or BIC.

The Akaike information criterion (AIC) (equation (16)) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a mean for model selection. The model with the lowest AIC is preferred.

$$AIC = 2(k+1) - 2\ln\left(\hat{L}\right), \tag{16}$$

where $\hat{L}$ is the maximum value of the likelihood function of the model to be tested and $k+1$ is the number of parameters to be estimated (k corresponding to the regressors plus one corresponding to the intercept).

AICc is AIC with a correction for finite sample sizes. If the assumption of a univariate linear model with normal residuals does not hold, then the formula for AICc will generally change. For more details see [6]. AICc is essentially AIC with a greater penalty for extra parameters. Using AIC, instead of AICc, increases the probability of selecting models that have too many parameters, i.e. of overfitting.

The Bayesian information criterion (BIC) or Schwarz criterion is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. The BIC can be calculated as:

$$BIC = \ln(N)(k+1) - 2\ln\left(\hat{L}\right), \tag{17}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model to be tested; N is the sample size; and $k+1$ is the number of parameters to be estimated.

3.4 Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning algorithm that models the relationship between a vector x of n input signals $x_i, i = 1, \ldots, n$ and an output signal y, that can be used both for classification and forecasting. It uses concepts borrowed from our understanding of how a biological brain responds to stimuli from sensory inputs. While the brain uses a network of interconnected cells, called neurons, to create a massive parallel processor, the ANN uses a network of artificial neurons or nodes to solve learning problems.

A typical artificial neuron with n input variables (dendrites), x_i, can be represented by:

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right), \tag{18}$$

where the weights w_i allow that each of the n inputs x_i contribute differently to the sum of input signals, and the resulting signal $y(x)$ is computed using an activation function $f(x)$. The roll of the activation function is to transform a neuron's net input signal into a single output signal to be broadcasted further in the network. Different activation functions have been proposed, such as, the unit step activation, linear, satured linear, hyperbolic tangent and gaussian [22]. However the most commonly used is the sigmoid activation function $f(x) = \dfrac{1}{1+e^{-x}}$, whose output values vary from 0 to 1. For this reason the sigmoid is sometimes called squashing function and it is necessary to standardize or normalize the input values in a small range around 0.

Choosing the activation function is an important element in the definition of the neural network. However, the network topology (or architecture) and the training algorithm are also important aspects.

The network topology, or architecture, describes the number of neurons in the model as well as the number of layers (i.e. groups of neurons) and how they are connected. In a single-layer network, there is a single set of connection weights

w_i and the input nodes process the incoming data exactly as it is received. On the other hand, multilayer network have one or more hidden layers that process the signals from the input nodes prior to reaching the output node. In terms of the direction of the flow of information through the network, one may have feedforward networks, when the input signal is fed continuously in one direction until the output, or recurrent network, whenever the signals are allowed to travel. The number of neurons in the hidden layers depends on many factors, such as, the number of input nodes, the amount of data for training the network, and the complexity of the task. Although there is no reliable rule to determine the number of neurons, a best practice consists on using the fewest nodes that result in adequate performance on a validation dataset.

The training algorithm specifies how connection weights are set in order to inhibit or excite neurons proportionally to the input signal. The backpropagation algorithm uses a strategy of back-propagating errors through the network in order to adjust the connection weights, and comprises a forward and a backward phase that performed in an iterative procedure until a stopping criterion is reached. It starts by randomly assigning values to the weight used and a gradient descent technique to determine the change of these weights in the iterative process. Although computationally expensive, this algorithm is commonly used due the good results typically obtained.

ANN can be adapted to classification or prediction problems and are among the most accurate modeling approaches. Furthermore, few assumptions about the data's underlying relationships are made. However, it results in a complex black box model that is difficult to interpret.

In this work, the sigmoid activation function, g, and a 0 to 1 range normalization defined by:

$$g(x) = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{19}$$

are used. For each forecast, the significant variables found in the regression analysis are used as predictors of the neural network. For evaluating the model performance, MSE, AIC and BIC are used.

4 Dataset and exploratory analysis

ALPHA is a Portuguese company which started its activity in 2006 focusing in the production and distribution of biofuel. Since then, the company has been continuously growing, extending its business areas to other fuels, now operating at a national level. In 2012 ALPHA started its LPG cylinder business. The company currently commercializes propane gas and has two types of cylinders, henceforth named A and B, with different capacity.

The dataset provided by the company included information about sales, return number of cylinders, operational stock of assets, total number of assets in the market, of the two types of propane cylinders (A – small bottles and B – large bottles), between January Year 0 (Jan/Y0) and December Year 1 (Dec/Y1). The forecast for sales during Year 2 was also provided by the company. Note that, for confidentiality reasons, the data used in this work was masked (with a conversion factor).

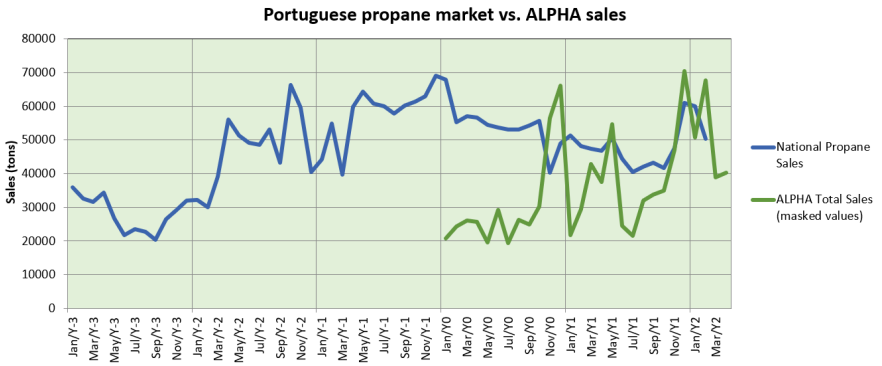

Fig. 2 Comparison between national sales and ALPHA total LPG sales.

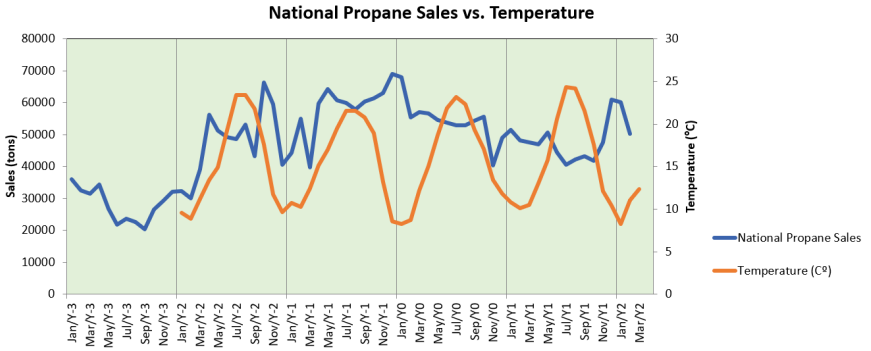

Fig. 3 Average temperature and national propane sales.

Additional data about consumption of propane in Portugal were also collected from the Portuguese Association of Petroleum Companies – APETRO[1]. Meteorological data such as temperature, humidity, precipitation, and wind were collected from OGIMET[2] and from the Portuguese Institute for Sea and Atmosphere – IPMA[3], respectively.

Figure 2 shows a general increase in the sales of ALPHA, however, it still represents a small percentage of the national market. Furthermore, the national and ALPHA sales of propane have different behaviours, thus in principle this is not a good indicator to forecast the company's sales when based on the national ones.

Figure 3 depicts the values of propane national sales and, simultaneously, the average temperature[4] in Portugal in the same period is shown. Since propane gas is used mostly for cooking and water heating, it is expected that whenever

[1] `http://www.apetro.pt`

[2] `http://www.ogimet.com`

[3] `https://www.ipma.pt`

[4] The air temperature in ($^{\circ}C$) was multiplied by a constant factor for a better comparison with the sales.

temperature decreases there is an increase in gas consumption. However, from Figure 3, this is not always true. In fact, for January Year -1 there is a decrease of the temperature and also a decrease of the consumption of gas.

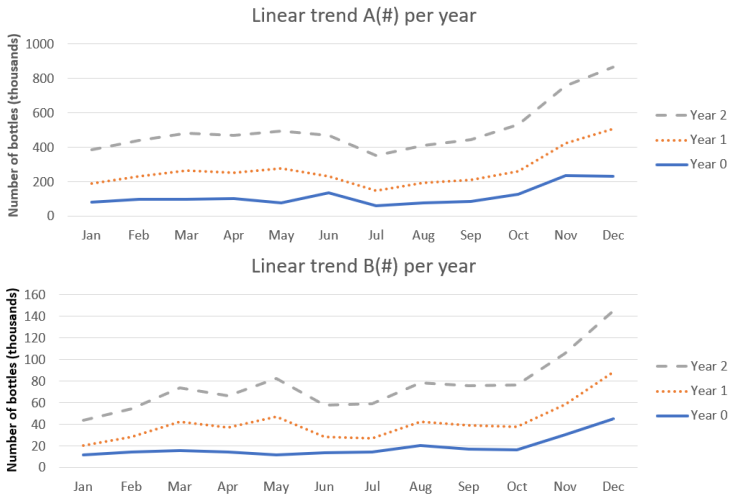

Fig. 4 ALPHA sales of type A and B cylinders across the years.

In order to study the existence of seasonality in sales of type A and B cylinders, Figure 4 presents the sales for the three years. In this figure, an increasing linear trend of ALPHA's sales, due to the company's market expansion, is observed. There are also some indicators of seasonality, since over the years the variations appear to be similar. In order to analyse the seasonality, moving averages seasonality coefficients and exponential smoothing forecasts are calculated in Section 5.

5 Bottled propane gas sales forecast

In order to make a good acquisition plan, forecasting sales is of the foremost importance. This Section is devoted to understand the variation of LPG sales, exploring several methods to forecast future sales of bottled propane gas and present the combination of the forecasts obtained by these methods. Forecasts for total propane sales in Portugal, total propane sales of ALPHA (in tons), sales of type A and B ALPHA assets are computed.

5.1 Time series models

5.1.1 Estimating seasonality coefficients

To study the seasonality of the data, seasonality coefficients for total propane sales of ALPHA (in tons); sales of type A and B ALPHA assets; and sales of butane, propane and total in Portugal were calculated.

Table 1 Seasonal coefficients of LPG sales

	PT			ALPHA		
Month	butane	propane	total	total	type A	type B
January	100%	99%	99%	84%	96%	69%
February	102%	95%	97%	109%	110%	108%
March	103%	93%	96%	96%	95%	98%
April	107%	111%	110%	89%	92%	84%
May	111%	110%	110%	125%	121%	132%
June	96%	98%	98%	82%	91%	70%
July	109%	98%	101%	59%	55%	65%
August	101%	98%	99%	81%	69%	97%
September	90%	92%	91%	80%	74%	87%
October	84%	109%	102%	87%	91%	81%
November	88%	99%	96%	135%	144%	124%
December	108%	98%	101%	172%	163%	183%

Let us consider, without loss of generality, ALPHA's sales of propane between January Year 0 and December Year 2. According with the subsection 3.1, the number of observations is given by $N = k \times s$, where k is the number of years and s the number of periods in the year, i.e. months. In this case $N = 3 \times 12 = 36$.

Using formulas (3.1), (2) and (3) it is possible to estimate the seasonal indexes of the sales presented in Table 1.

The national coefficients (PT butane, PT propane, PT total), presented in Table 1, do not have significant variations from month to month. In fact, for the national sales, the PT butane seasonal coefficients present a minimum of 84% in October and a maximum of 111% in May, while for PT propane and PT total the minimum is attained in September with 92% and 91%, respectively, and the maximum occurs in April with 111% and 110%.

On the other hand, ALPHA's sales show more significant variations. The seasonal coefficients of the total sales vary from 59% in July to 172% in December, while for type A cylinders from 55% in July until 163% in December, and finally, type B cylinder sales vary from 65% in July to 183% in December (see Table 1). The months with high coefficients (above 100%) are February, May, November and December, and the ones with smallest coefficients (below 100%) are June, July, August and September, corresponding to summer season. Therefore, forecast for sales can be done using these seasonal coefficients.

In Figure 6, the seasonally adjusted series using these coefficients is plotted in dashed red. A linear trend line (in thin red) was added to the chart, in order to perceive ALPHA's sales growth. It was found that the company is increasing their market share, and sales are growing at a rate of 56.8 tons per month.

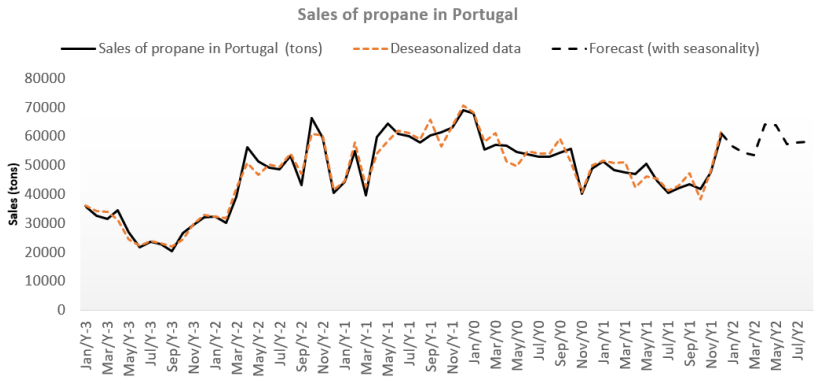

Fig. 5 Forecast of propane national sales using Holt's method (in dashed black), together with the seasonally adjusted series (in dashed red).

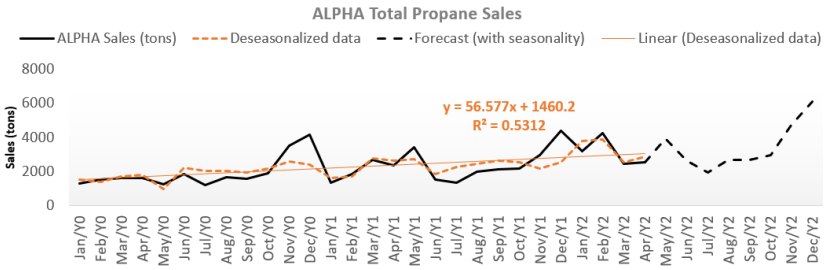

Fig. 6 Forecast of ALPHA total sales using Holt's method (in dashed black), together with the seasonally adjusted series (in dashed red).

5.1.2 Exponential smoothing forecast

The data concerned with bottled propane gas sales was deseasonalized using the seasonal coefficients, obtained in the previous subsection 5.1.1, then Holt's method was used to forecast sales.

The level coefficients α and trend coefficients β obtained for the models using Holt's method [38] applied to the seasonally adjusted series are presented in Table 2). The error measures regarding the training set in Holt's method were also computed and presented in this table. After obtaining the estimated values of Holt's method for the deseasonalized data, the re-composed series was obtained by multiplying the estimated values (and forecasts) by the seasonal coefficients. These final estimated values for the time series were used to compute the final MSE presented in the last line of the table.

In Figures 5 and 6 the forecast for the company's total sales are compared with the forecast for the national sales of propane gas. These forecasts result from the same time series decomposition method used with the type A and B bottles series.

In Figure 7, the observed values of the type A and B bottles sales (from Jan/Y0 to Apr/Y2) are presented alongside with the estimated values (from Jan/Y0 to Apr/Y2), and the forecast sales for May/Y2 until Dec/Y3. The forecast for the

Table 2 Details of the models obtained using Holt's method for forecasting sales of National Propane, total ALPHA sales, and type A and B bottles.

	National tons	ALPHA tons	Type A bottles	Type B bottles
α	0.6321	0.0001	0.0253	0.0001
β	0.0001	0.0001	0.0253	0.0001
AIC	1309.326	442.8589	671.9332	584.8535
AiCc	1310.437	445.5862	674.6605	587.5808
BIC	1319.798	449.5199	678.9542	591.5146
RMSE	6506.423	429.8864	25697.11	5426.988
MAE	4891.788	348.4303	18961.96	4040.577
MAPE	10.80773	16.35543	14.93226	20.39841
MSE	42333540.25	184802.3169	660341462.4	29452198.75
Final MSE	43410331.42	211117.2891	771524445.6	30358125.74

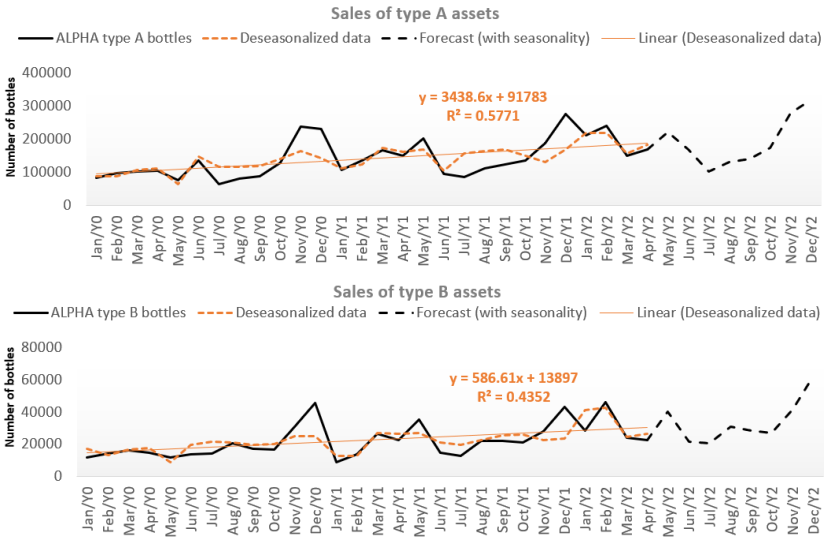

Fig. 7 Forecast of type A and B bottle sales using Holt's method (in dashed black), together with the seasonally adjusted series (in dashed red).

type A and B bottles series are compared. These forecasts are the result of the same time series decomposition method used with the national sales of propane gas and company's total sales. These series were first deseasonalized with the seasonal indexes computed in Table 1 and then Holt's method was applied. Finally, the forecasts were multiplied by the seasonal coefficients to reflect the monthly fluctuations. The details of these models are summarized in Table 2.

5.2 Linear Regression Models

Linear Regression models were also explored in this challenge. Data provided by ALPHA, concerning sales from Jan/Y0 to Apr/Y2, was used to estimate the regression models for ALPHA's sales of propane, and sales of type A and B ALPHA assets. Subsequently these models were used to forecast the sales for the period from May/Y2 to Dec/Y2. Data regarding total propane sales in Portugal from Jan/Y0 to Jan/Y2 was used for the estimation of the corresponding regression

model, that was subsequently used to forecast the national sales for the period from Jan/Y2 to Aug/Y2.

In order to apply the MLR approach, it is important to determine which variables may influence gas sales. The following variables were used as predictors in the previously mentioned models:

- `Temperature`: Temperature, in °C;
- `PromoCampaign` – ALPHA's promotional campaign – 0 if no promotional campaign or 1 if there is a promotional campaign ongoing;
- `SalesObjective` – ALPHA's sales objective – 0 for no or 1 for yes;
- `ExpectPriceIncrease` – ALPHA's expectation of price increase – 0 for no or 1 for yes;
- `Wind` – the monthly mean wind velocity (km/h);
- `Month` – being Month= 1 for January, ..., Month= 12 for December.
- `Time` – elapsed months since Jan/Y0;

However, after testing several combinations of these variables, not all of them were found to be relevant in all cases. Wind conditions were not significant in any of the experimented models. Also, the month of the year was not found to be a good factor to be used in the models. Instead, elapsed time was considered in all models, for being able to predict future values of the series.

Table 3 Summary of the multiple linear regression models obtained for national sales of propane, ALPHA total propane sales, and ALPHA sales of type A and B bottles.

	Propane National Sales (tons)			ALPHA Total Sales (tons)			Sales of Type A bottles			Sales of Type B bottles		
	β	SE	p-value	β	SE	p-value	β	SE	p-value	β	SE	p-value
(Intercept)	65195.0383	3911.9	2.47e-14	1978.519	252.403	8.26e-08	140068.9	13086.1	3.45e-10	11821.2	2021.9	5.87e-06
Temperature	-561.9096	206.4	0.01215	-51.224	13.541	0.001022	-4351.5	702.0	3.08e-06			
PromoCampaign				1228.198	215.97	1.02e-05	86935.7	11197.2	9.67e-08	9527	3065.8	0.004958
SalesObjective				1919.109	266.171	3.17e-07	99561.8	13800.0	3.14e-07	23969.1	3687.7	1.24e-06
ExpectPriceIncrease				1458.265	375.292	0.000797	59204.7	19457.4	0.00597	21849.3	5267.9	3.90e-04
Time	-406.769	144.1	0.00964	52.17	8.591	4.11e-06	3439.3	445.4	1.06e-07	467.1	122.4	0.000886
df	23			22			22			23		
R^2	0.3977			0.8921			0.921			0.8003		
R^2_{adj}	0.3453			0.8676			0.903			0.7655		
F	7.593		2.94e-03	36.4		6.34e-10	51.29		2.157e-11	23.04		9.21e-08
Shapiro-Wilk	0.9709		0.6468	0.95596		0.2786	0.98115		0.877	0.95864		0.3236
Breusch-Pagan	4.5498		0.1028	3.4005		0.6385	9.322		0.09689	5.985		0.2003
Breusch-Godfrey	3.3456		0.06739	0.4938		0.4822	2.7107		0.09968	0.069881		0.7915
max VIF	1.000168			1.125327			1.125327			1.116417		
MSE	26851317			94826.64			254896358			20135026		

5.2.1 National Sales of Propane

For the national propane sales series, only the temperature was found to be a significant predictor, as may be seen in Table 3. As expected, the company's related variables such as the existence of promotional campaigns, sales objectives, and expectation of price increase do not affect the national sales of LPG.

$$\begin{aligned} NationalSales = 65195.04 - 561.91\,Temperature \\ -406.77\,Time + \varepsilon, \qquad \varepsilon \sim N(0,\sigma^2)\ i.i.d. \end{aligned} \tag{20}$$

The negative coefficient of the variable `Time` in equation (20) means that propane national sales are generally decreasing at an average rate of 406 tons

per month, when other variables remain constant. Also, for each Celsius degree increase in the temperature, the national sales would decrease in average nearly 562 tons (considering the same elapsed time) and other variables constant. The adjusted determination coefficient of this model is $R^2_{adj} = 0.3453$, revealing that this linear regression model is weak, and that temperature and time explain just a small part (34.5%) of the variability of the national sales of propane. Even though, the F test and the residual analysis confirms it is a valid regression model.

5.2.2 ALPHA's Total Sales of Propane

The best adjusted model, equation (21), for the company's total sales, whose details are shown in Table 3, has $R^2_{adj} = 0.8676$, therefore 86.7% of the variability of ALPHA's total sales of propane gas, in tons, is explained by the variables `Temperature`, `PromoCampaign`, `SalesObjective`, `ExpectPriceIncrease`, and `Time`, which are all significant at a 5% level. The obtained coefficients show that the company's sales have generally been increasing, at an average rate of 52 tons per month, and that promotional campaigns have significant effect of increasing the sales in 1228 tons (considering all other variables constant). The months when the company's distributors have to meet sales objectives also have a significant impact on sales, increasing them in almost 2 thousand tons. In the months where there is an expectation of price increase, sales are to increase 1458 tons. Variable temperature also has a negative impact on the company's total sales. Considering all other variables constant, for each Celsius degree increased, ALPHA sales decrease 51 tons; fortunately, this also means that for each degree that the temperature drops, sales will increase 51 tons.

$$\begin{aligned} ALPHA\_TotalSales = {} & 1978.52 - 51.22\,Temperature \\ & +1228.20\,PromoCampaign \\ & +1919.11\,SalesObjective \\ & +1458.27\,ExpectPriceIncrease \\ & +52.17\,Time + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)\ i.i.d. \end{aligned} \tag{21}$$

A residual analysis was performed to this model. From Table 3, Shapiro-Wilk test resulted in a p-value of $0.2786 > 0.05$, confirming that the errors can be assumed to be normally distributed. The p-value of the Breusch-Pagan test was $0.6385 > 0.05$, not rejecting homoscedasticity, and Breusch-Godfrey test p-value was $0.4822 > 0.05$, confirming the independence of the residuals. All the predictor variables in the model produced VIF values below 1.125, stating that there are not any multicollinearity problems.

5.2.3 Type A Bottles Sales

In terms of the number of type A bottles sales, different regression models were tested. equation (22), where `ALPHA_A_Sales` represents the dependent variable,

describes the best model found.

$$\begin{aligned} ALPHA\_A\_Sales = 140068.9 &- 4351.5\, Temperature \\ &+ 86935.7\, PromoCampaign \\ &+ 99561.8\, SalesObjective \\ &+ 59204.7\, ExpectPriceIncrease \\ &+ 3439.3\, Time + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)\ i.i.d. \end{aligned} \tag{22}$$

The adjusted coefficient of determination was $R^2_{adj} = 0.903$, that is 90.3% of the variability of the number of bottles of type A sold is explained by the variables that are included in the model.

The normality assumption and the homoscedasticity assumption of the errors in the model were evaluated and validated, through the application of the Shapiro-Wilk test and the Breusch-Pagan test which the results are presented in Table 3.

In the right side of Figure 8, the autocorrelation function of the residuals shows small values for all lags, which is corroborated by the p-value of the Breusch-Godfrey test $0.09968 > 0.05$, thus confirming independence of the residuals. Also, from the fair dispersion in the residuals plot on the left side of Figure 8, the model in equation (22) can be considered adequate. Furthermore, all VIF values were smaller than 3 (the maximum was 1.125), for which no multicollinearity exists.

Having into account the linearity of model in equation (22), one can interpret each one of its parameters, assuming the others variables fixed. Therefore:

- an increase of 1°C of mean temperature will result in a decrease of the sales of type A bottles by approximately 4352 units;
- when ALPHA has a promotion campaign, sales of type A bottles increase in approximately 86936 units;
- when sales objectives are set, there is an increase of sales of approximately 99562 units;
- whenever there is expectation of prices increase there is an increase of sales of approximately 59205 units;
- for every month that passes after January Y0, there is an increase on the sales of approximately 3439 units.

Figure 9 presents, in black, the observed number of type A propane gas bottles from January Year0 to May Year2 and the company's estimates from June to December Year2, while in red are presented the estimates from January Year0 to May Year2 obtained with model in equation (22). The forecasts were computed using similar temperature conditions as the previous years.

The existence of promotional campaigns presents a relatively high impact on the sales. For this reason, two different scenarios are considered. In scenario 1 the existence of the company's promotional campaigns in June and November is considered, while in scenario 2 it is assumed that no promotional campaigns occur.

The estimated values of the type A bottles sales closely replicate the behavior of the past observed sales. For both scenarios, the forecast obtained using the regression model in equation (22) presents lower values than the ones estimated by the company, therefore the company's expectation of growth is larger than the one predicted by the proposed regression model (Figure 9 and Table 4). Since the authors do not have information concerning the company's growth strategy,

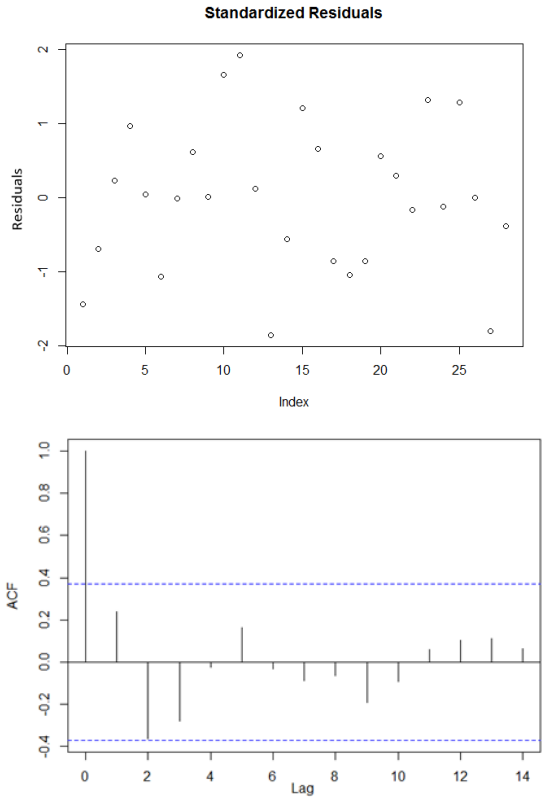

Fig. 8 Analysis of the residuals of the model for ALPHA's type A bottle sales (Eq 22).

perhaps the company is aiming at further promotional campaigns or other actions that are expected to increase the volume of sales.

Table 4 Predicted values obtained by the regression model in equation (22) for type A assets sales, considering scenario 1 and 2, and the company's sales expectation.

Year	Month	MLR Model Forecast under Scenario 1	MLR Model Forecast under Scenario 2	Company's Sales Expectations
Year 2	May	167 693	167 693	216 545
	June	241 891	154 956	239 624
	July	146 102	146 102	204 792
	August	150 640	150 640	218 204
	September	162 935	162 935	235 039
	October	179 526	179 526	270 725
	November	292 453	205 517	333 270
	December	319 451	319 451	360 066

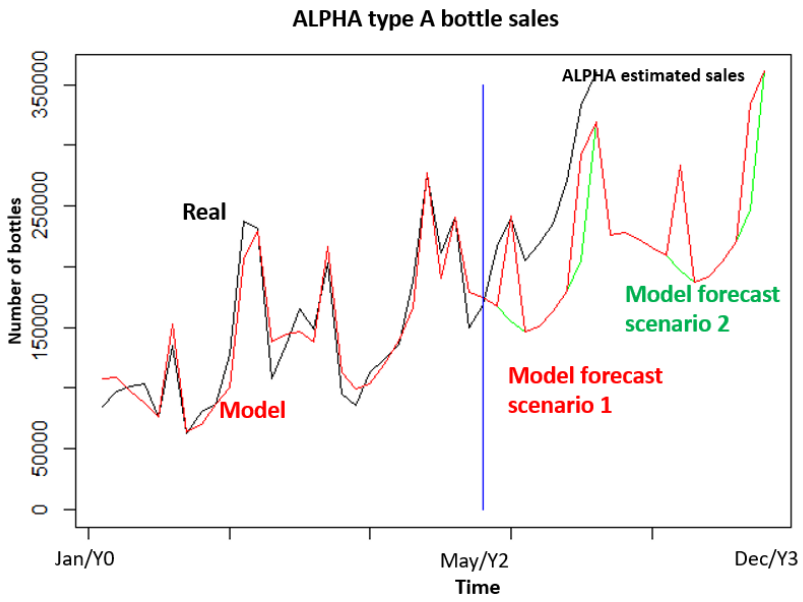

Fig. 9 Observed and estimated sales, and forecast of the ALPHA's type A bottles sales, using the model from equation (22), for scenarios 1 and 2. See text for further details.

5.2.4 Type B Bottles Sales

In terms of the ALPHA's type B bottles sales, the best model found is given by equation (23), where `ALPHA_B_Sales` represents the number of assets of type B sold by ALPHA.

$$\begin{aligned} ALPHA\_B\_Sales = {} & 11821.9 + 9527.0\, PromoCampaign \\ & +23969.1\, SalesObjective \\ & +21849.3\, ExpectPriceIncrease \\ & +467.1\, Time + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)\ i.i.d. \end{aligned} \tag{23}$$

From Table 3 one can see that 76.6% of the variability of the number of bottles of type B sold is explained by the variables included in model of equation (23). In this model, the temperature revealed not to be significant, which can be due to the fact that these kind of bottles are used mostly in industries and not so dependent on the temperature of the month. The analysis of residuals was also performed. The normal distribution of the residuals was not rejected through the application of the Shapiro-Wilk test ($W = 0.958$, p-value=0.324). With a p-value of 0.2003 in the Breusch-Pagan test, homoscedascity can also be assumed; independence of the residuals is also confirmed by the p-value of 0.7915 in the Breusch-Godfrey test. Furthermore, all VIF values were smaller than 1.12, for which no multicollinearity problems exist in the model.

The model in equation (23) can be used to forecast sales for assets B, considering several scenarios that have into account the existence or not of promotional

campaigns, sales objectives and also the expectation of price increase. The forecasts considering scenario 1, confronted with the company's sales expectations, are presented in Table 5.

Table 5 Predicted values obtained by the regression model in equation (23) for type B assets sales, considering scenario 1, and the company's sales expectation.

Year	Month	MLR Model Forecast under Scenario1	Company's Sales Expectations
Year 2	May	25 366.95	35 329
	June	35 361.10	29 540
	July	26 301.14	32 025
	August	26 768.24	36 001
	September	27 235.33	36 736
	October	27 702.43	38 549
	November	37 696.57	47 292
	December	52 605.70	56 784

5.3 Artificial Neural Networks

For the neural network models, the same predictors as for the regression models were used. Relatively to ALPHA's sales of propane, sales of type A and B assets, data from Jan/Y0 to Apr/Y2 was used for training the networks, and sales from May/Y2 to Dec/Y2 were forecasted. For total propane sales in Portugal it was used data from Jan/Y0 to Dec/Y1 for training, and sales from Jan/Y2 to Aug/Y2 were forecasted.

The networks were trained using the package *neuralnet* of **R** software. The logistic activation function and the resilient backpropagation algorithm with weight backtracking, with a threshold of 0.01 was used. Single layer neural networks with up to four hidden nodes were ran 100 times each.

The best neural network was chosen in terms of the smallest MSE. These ANN, with their corresponding weights, for propane national sales, total ALPHA sales, and type A and B bottles are depicted in Figures 10 and 21. In this figure, error stands for sum of squared errors, $SSE = n \times MSE$. Table 6 shows the results of the best of the 100 runs for the propane national sales, total ALPHA sales, and type A and B bottles. Note that the AIC and BIC error measures (equations (16) and (17)) with this method are considerably lower than the values obtained in Table 12 for the Holt's method.

Table 6 Summary of the ANN obtained for sales of national propane, total ALPHA sales, and type A and B bottles.

	National sales	ALPHA sales	Type A	Type B
# nodes	4	3	3	3
# steps	1109	361	241	735
AIC	34.132	44.160	44.084	38.247
BIC	54.159	73.468	73.393	3.559
MSE	2 105 184.116	34 024.6106	7 848 7460.700	6 996 131.238

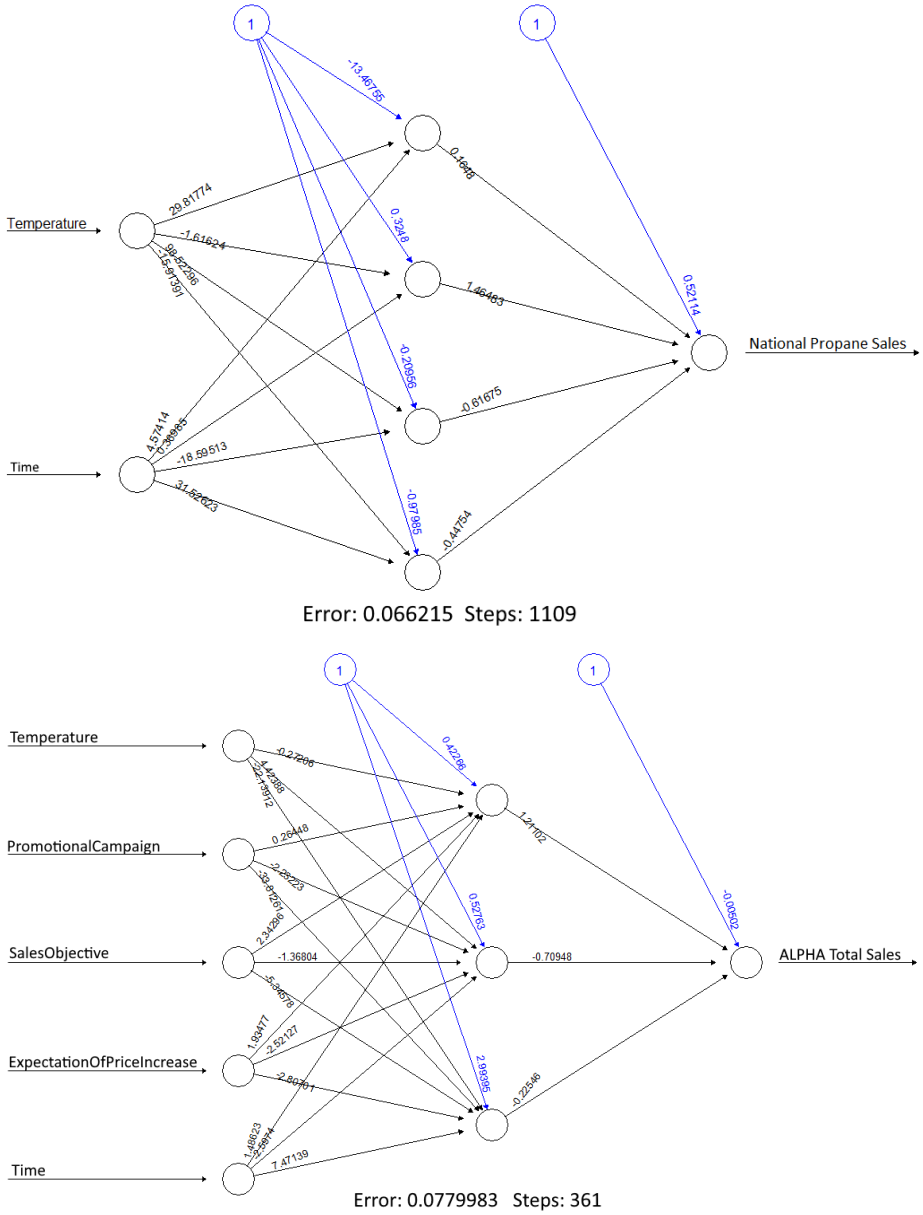

Fig. 10 The best ANN for propane national sales and total ALPHA sales.

For national sales, only two predictors were considered and therefore the network presents only two input nodes. The best ANN has four hidden nodes and a total of 1109 steps were computed in order to obtain a MSE of 2105184.116, a value that is significantly smaller than the one obtained using TS and MLR models, more precisely 5% and 8%, respectively.

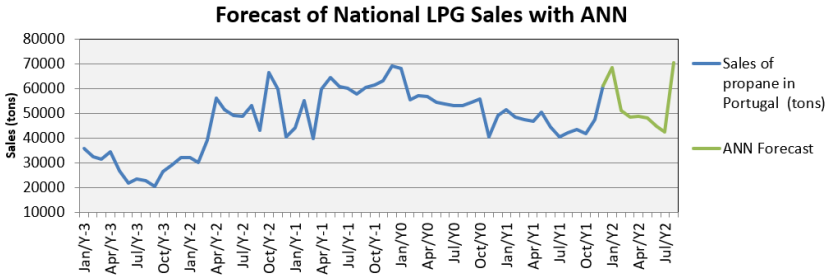

Fig. 11 Forecasts obtained for total propane sales in Portugal (from Jan/Y2 to Aug/Y2) using neural networks.

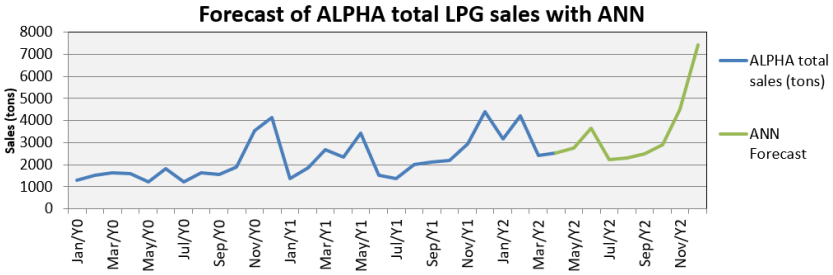

Fig. 12 Forecasts obtained for ALPHA's total sales of propane (in tons), (from May/Y2 to Dec/Y2) using neural networks.

The neural network used for ALPHA total sales presents five input nodes, correspondent to the five predictors. The best neural network has three hidden nodes and a MSE of 34024.6106, which is smaller than the one obtained using TS and MLR models, 16% and 36%.

For type A and type B bottles sales, the networks have five and four input nodes, respectively. In both, the best ANN presents three hidden nodes, and has a MSE considerably smaller than the ones obtained using TS and MLR methods. In fact, the MSE obtained using ANN is around 30% of the MSE obtained using MLR (both for type A and type B bottles sales), and is 10% of the MSE obtained using TS for type A bottles and 20% for type B bottles.

The best network for the total propane sales in Portugal (see Figure 10 and Table 6) was used to obtain the forecast of the sales for period from Jan/Y2 to Aug/Y2. The resulting forecast is shown in Figure 11. The forecasts for the period from May/Y2 to Dec/Y2, of ALPHA's total sales of propane, sales of type A and B assets were obtained using the best models. The resulting forecast are shown in Figure 12. They resemble the behavior of the ALPHA sales in previous year and the general increase in sales indicating the growth of the company.

5.4 Combining the forecasts

Using TS, MLR and ANN techniques it was possible to obtain punctual forecasts for each month, m. However, these forecasts have inherent errors. In order to take into account the errors inherent to each forecast technique, it was

Table 7 Weight assigned to each forecasting method for the national sales of propane, and total ALPHA sales, and type A and B bottles sales.

SALES	α_{TS}	α_{LM}	α_{ANN}
Propane national sales	4.30%	6.96%	88.74%
ALPHA total sales	10.60%	23.61%	65.79%
Type A bottles	7.22%	21.84%	70.94%
Type B bottles	14.60%	22.02%	63.37%

considered that the forecast of each algorithm, for each month m, follows a normal probability distribution with mean equal to the punctual value of the forecast, $\bar{Y}_{TS}(m)$, $\bar{Y}_{MLR}(m)$, $\bar{Y}_{ANN}(m)$, and variance equal to the MSE, MSE_{TS}, MSE_{MLR}, MSE_{ANN}, i.e.:

$$Y_{TS}(m) \sim \mathcal{N}(\bar{Y}_{TS}(m), MSE_{TS}), \tag{24}$$

$$Y_{MLR}(m) \sim \mathcal{N}(\bar{Y}_{MLR}(m), MSE_{MLR}), \tag{25}$$

$$Y_{ANN}(m) \sim \mathcal{N}(\bar{Y}_{ANN}(m), MSE_{ANN}). \tag{26}$$

For each month, m, the probability distribution of the sales forecast obtained by each forecasting method (TS, MLR and ANN) are combined linearly. Since the errors of each technique are different, the magnitude of these errors is used to determine the percentual contribution of each forecasting for computing a forecast that combines all the forecasts obtained by the different techniques. Therefore, standard weights, inversely proportional to the magnitude of the forecast error, are considered. More precisely, the weights are computed such that:

$$\sum_{i=1}^{3} \alpha_i = 1,\ i = 1, 2, 3$$

and

$$\alpha_i\, MSE_i = \alpha_j\, MSE_j,\ i \neq j = 1, 2, 3.$$

With this multi-model methodology, the methods that have a larger contribution to the combined forecast are the methods with minor MSE.

Table 7 presents the weights assigned to each forecasting methods. For propane national sales, the ANN forecast contributes with 88.74% to the combined forecast, while MLR contributes with 6.96% and finally TS presents the smallest contribution (4.30%). For ALPHA total sales, type A bottles, type B bottles sales, the ANN also presents the highest contribution, while TS present the lowest contribution.

Now that the weights (α_{TS}, α_{LM} and α_{ANN}) have been computed, the combined forecast for each month $m = 1, \ldots, 8$ is given by the weighted linear combination of the forecast obtained using each forecasting techniques, as presented in equation (1). For this purpose, five Monte Carlo simulation runs were used, with a sample size of 150.

Table 8 presents the results for the combined forecasts of national sales for the months from Jan/Y2 to Aug/Y2. August is the month that presents the highest average sale forecast and the smallest standard deviation. The forecast for January presents averages quite close to the sales for August, but with a higher variation. This behavior on sales resembles the one observed in previous years, since the month of January was the one that presented the second highest sales on Y1.

Table 8 National sales forecast

Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval	
			Lower bound	Upper bound
Jan/Y2	66 695.52	106.69	66 601.99	66 789.04
Feb/Y2	50 911.95	93.270	50 830.20	50 993.70
Mar/Y2	48 443.33	132.98	48 326.77	48 559.90
Apr/Y2	49 411.81	134.41	49 293.99	49 529.63
May/Y2	48 635.16	132.76	48 518.79	48 751.53
Jun/Y2	44 955.04	97.360	44 869.70	45 040.39
Jul/Y2	42 769.19	62.970	42 713.99	42 824.38
Aug/Y2	67 518.91	59.340	67 466.90	67 570.92

Table 9 ALPHA sales forecast

Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval	
			Lower bound	Upper bound
May/Y2	2 845.32	13.62	2 833.38	2 857.26
Jun/Y2	3 550.21	6.56	3 544.46	3 555.96
Jul/Y2	2 261.07	10.77	2 251.63	2 270.51
Aug/Y2	2 379.93	14.32	2 367.38	2 392.49
Sep/Y2	2 546.78	13.90	2 534.60	2 558.96
Oct/Y2	2 891.00	13.55	2 879.12	2 902.88
Nov/Y2	4 512.62	5.12	4 508.13	4 517.11
Dec/Y2	6 775.04	4.96	6 770.69	6 779.38

From March to April it is also expected to exist a slightly increase on the sales followed by a decrease in the following month. In fact, from the historical data the months from March to May present a slight increase in sales with values very close for these three months.

In Figure 13 are depicted in blue the observed data until Dec/Y1, and in green the average response of the combined forecast of national sales from Jan to Aug/Y2.

Table 9 presents the results for the combined forecasts of ALPHA sales for the months from May to Dec/Y2. December is the month with the highest average sales forecast and the smaller variation, which is in line with behavior observed in ALPHA's sales in previous years. It is also expected that the sales slightly increase from May to June flowed by a small decrease in July. The highest variations in May and July than in June suggest that the increase can also take place in May

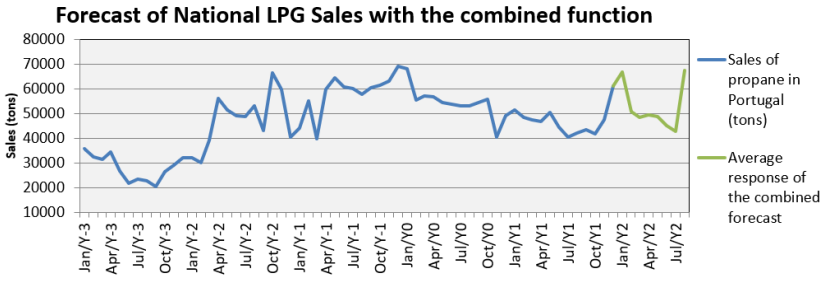

Fig. 13 Combined forecasts for national sales of propane.

Table 10 ALPHA sales type A and type B bottles forecast

	Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval Lower bound	95% Confidence Interval Upper bound
SalesA	May/Y2	166 193.40	764.49	165 523.30	166 863.50
SalesA	Jun/Y2	177 281.30	453.48	176 883.80	177 678.80
SalesA	Jul/Y2	134 030.10	523.45	133 571.30	134 488.90
SalesA	Aug/Y2	140 075.00	761.46	139 407.60	140 742.50
SalesA	Sep/Y2	152 109.80	674.65	151 518.50	152 701.20
SalesA	Oct/Y2	168 441.50	532.39	167 974.80	168 908.10
SalesA	Nov/Y2	260 213.70	772.22	259 536.80	260 890.50
SalesA	Dec/Y2	442 973.00	993.14	442 102.50	443 843.50
	Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval Lower bound	95% Confidence Interval Upper bound
SalesB	May/Y2	27 125.81	155.74	26 989.29	27 262.32
SalesB	Jun/Y2	37 153.10	180.43	36 994.95	37 311.26
SalesB	Jul/Y2	24 981.13	191.91	24 812.92	25 149.35
SalesB	Aug/Y2	26 560.19	108.08	26 465.46	26 654.93
SalesB	Sep/Y2	26 649.41	127.96	26 537.25	26 761.57
SalesB	Oct/Y2	26 691.51	146.63	26 562.98	26 820.04
SalesB	Nov/Y2	41 498.01	194.58	41 327.46	41 668.57
SalesB	Dec/Y2	71 265.38	183.96	71 104.13	71 426.63

or June. In Figure 14 are depicted in blue the observed data until Apr/Y2, and in green the average response of the sales from May to Dec/Y2.

Table 10 presents the results for the combined forecasts of sales of type A and B bottles, respectively. During the months from May to Dec/Y2, there is variation in the average sales of in approximately 309 000 type A bottles and 46 000 type B bottles. The average sales, for type A and type B bottles is highest in December and smaller in July, which is in line with the behavior observed in previously periods in ALPHA sales. It is also clear the generic increase in sales through the year that reflects the growth of ALPHA's market.

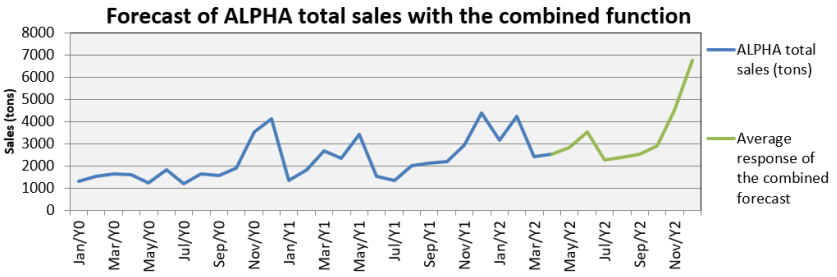

Fig. 14 Combined forecasts for total ALPHA total sales.

In Figure 14 are depicted in blue the observed data until Apr/Y2, and in green the average response of the sales from May to Dec/Y2.

6 Return rate forecast

To correctly plan the acquisition of new bottles from the supplier, not only demand must be known, but also the reverse logistic flows. The empty bottles being returned to ALPHA can be reinserted in the system, filled again and sold to the clients. As the acquisition of new bottles is expensive, reusing is the key. This Section is devoted to understand the fluctuation of returns and to forecast the return rate.

6.1 Time series models

Similar to Section 5, the seasonal coefficients for returns of type A and B bottles can be calculated. Figure 15 contains these values.

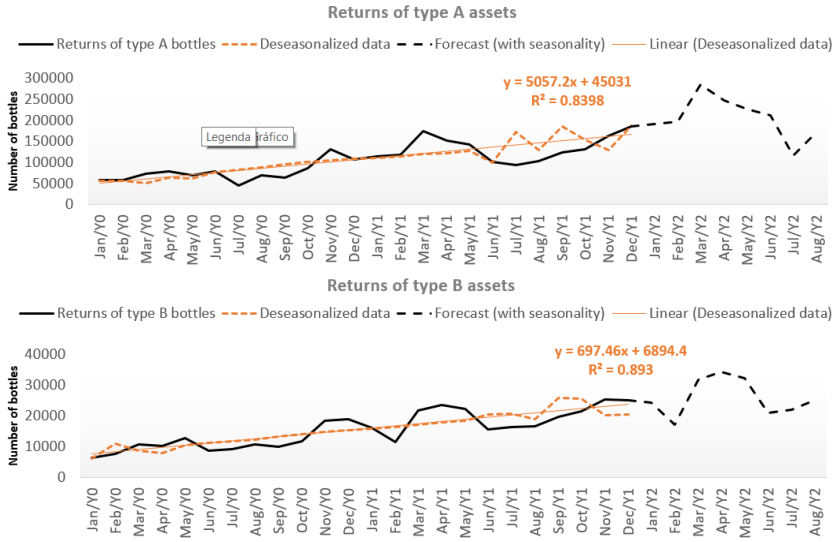

Fig. 15 Forecast of type A and B bottles returns using only the seasonal coefficients

The obtained seasonal coefficients are in Table 11. In the summer months the returned bottles are usually well below average, while in the beginning of Winter and in Easter the returned bottles have its peak.

In Figure 15, the seasonally adjusted series using these coefficients is plotted in dashed red. A linear trend line was added to the chart, in order to perceive the growth of the number of returns. The authors found that type A bottles returns are increasing at a rate of 5057.2 bottles per month, while type B returns are also increasing, but a lower rate of 697.5 bottles per month.

With an analogous methodology to the one for forecasting sales, the authors used time series decomposition and applied Holt's method to forecast the number of returned bottles of both types. The details of Holt's method are presented in Table 12. The final MSE was computed considering the forecast values after

Table 11 Seasonal coefficients for the returned bottles.

Month	typeA returns	typeB returns
January	102%	101%
February	103%	70%
March	146%	126%
April	125%	132%
May	112%	121%
June	102%	77%
July	55%	79%
August	79%	87%
September	67%	76%
October	86%	84%
November	125%	124%
December	99%	124%

recomposing back the series with the seasonal coefficients, which are plotted in dashed black in Figure 15.

Table 12 Details of the models obtained using Holt's method for forecasting returns of ALPHA's type A and B bottles.

	Type A bottles	Type B bottles
α	0.9218	0.0291
β	0.0001	0.0001
AIC	567.62	472.0148
AiCc	571.0253	475.3481
BIC	573.5822	477.9051
RMSE	22687.2	3091.096
MAE	17585.22	2719.817
MAPE	18.05655	19.7187
MSE	514709043.8	9554874.481
Final MSE	736459515.7	2928710.637

6.2 Linear Regression Models

The application of linear regression to estimate and predict the number of return of type A ($ALPHA\_A\_Returns$) and of type B ($ALPHA\_B\_Returns$) bottles allowed to obtain the best regression models described in equations (27) and (28), respectively.

$$\begin{aligned} ALPHA\_A\_Returns = {} & 100328.726 - 3688.228\, Temperature \\ & +25313.432\, PromoCampaign \\ & +4834.547\, Time + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)\ i.i.d. \end{aligned} \tag{27}$$

According to the results in Table 13, approximately 86.04% of the variability of the returns of type A bottles is explained by the temperature, the existence of promotional campaigns, and the elapsed time in months from January Y0. This model has a rather good explaining power.

Each of the coefficients in equation (27) can be interpreted, assuming all other variables fixed. Therefore:

– an increase of 1°C of mean temperature will result in a decrease of 3688 returned type A bottles;

Table 13 Summary of the multiple linear regression models obtained for the company's type A and B number of returned bottles.

	Returned Type A bottles			Returned Type B bottles		
	β	SE	p-value	β	SE	p-value
(Intercept)	100328.726	10509.7	6.87e-09	11129.2745	1948.81	1.14e-05
Temperature	-3688.228	589.2	4.12e-06	-309.3492	109.07	0.00989
PromoCampaign	25313.432	9068.9	0.0113			
Time	4834.547	437.6	5.77e-10	739.5246	81.08	9.46e-09
df	20			21		
R^2	0.8786			0.8022		
R^2_{adj}	0.8604			0.7833		
F	48.27		2.42e-09	42.57		4.09e-08
Shapiro-Wilk	0.9101		0.03545	0.96099		0.4587
Breusch-Pagan	3.3631		0.339	2.9506		0.2287
Breusch-Godfrey	1.7128		0.1906	0.65652		0.4178
max VIF	1.034489			1.026144		
MSE	177734850			6446543		

– when ALPHA has a promotion campaign, returns of type A bottles increase in approximately 25313 units;
– for every month that passes after January Y0, there is an increase on the number of returned type A bottles of approximately 4835 units.

The assumptions on the residuals were tested, and the small Shapiro-Wilk p-value of 0.03545 may mean that the residuals are not normally distributed, as they should. In Figure 16 on the left, two large residuals arise. The Breusch-Pagan test yielded a p-value of 0.339, confirming the homoscedasticy assumption, and both the Breusch-Godfrey test p-value of 0.1906 and the small values of the Autocorrelation Function of the residuals on the right side of Figure 16 confirm the independence. The maximum VIF is 1.0345 which clearly states that there is no multicollinearity problems.

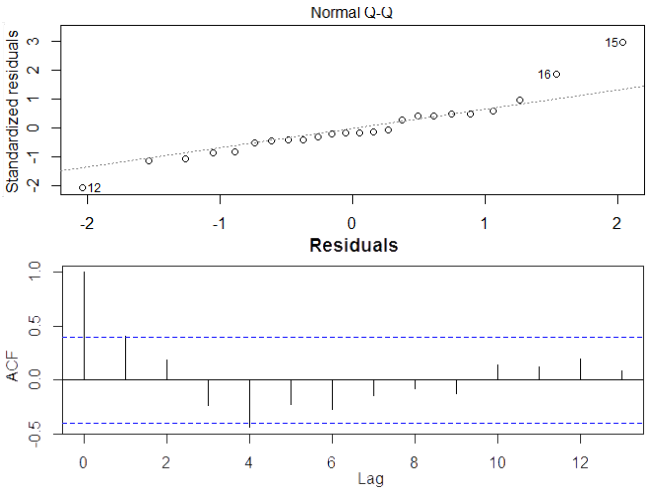

Fig. 16 Normal Q-Q plot and AutoCorrelation Function of the residuals of the linear regression model in equation (27) for the number of returned bottles of type A.

This model was used to forecast the number of returned type A bottles assuming temperature conditions equivalent to the previous year and no promotional campaigns, scenario 1. The forecast results are shown in Table 14.

Regarding the number of type B returned bottles, the obtained model does not reflect sensitivity to promotional campaigns. However, contrary to its sales, returns were found sensitive to the temperature, having a negative effect with an average of minus 309 bottles returned per degree Celsius increased.

$$ALPHA\_B\_Returns = 11129.2745 - 309.3492\,Temperature + 739.5246\,Time + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)\ i.i.d. \tag{28}$$

Approximately 78.3% of the variability of the number of returned bottles of type B is explained by the temperature and the elapsed time since January Y0. The residual analysis confirmed all the assumptions of the linear regression model. Therefore, it was used to predict the number of type B assets returned to the company between January and August in Year 2, as presented in Table 14.

Table 14 Predicted values obtained by the regression model in equation (27) and 28 for type A and B assets returns considering scenario 1.

	MLR Model Forecast for	
Year 2	Type A returns	Type B Returns
January	190727.6271	27062.16427
February	185272.0164	26938.60466
March	185274.9838	27272.88181
April	187527.7706	27795.86195
May	179407.4148	27448.79756
June	195844.4042	27038.3166
July	164946.2734	26903.92976
August	170712.0978	27721.56498

6.3 Artificial Neural Networks

For modeling the returns of type A bottles using ANN, Temperature, PromoCampaign and Time were used as predictors - the same as for the regression models - therefore the network presents 3 input nodes. While, for modeling the returns of type B bottles only Temperature and Time, were used as predictors, in accordance to what was done for regression models (see Figures 17 and 22).

For both type of bottles return it was used data from Jan/Y0 to Dec/Y1 for training, and data from Jan/Y2 to Aug/Y2 for forecast. The networks were trained using the logistic activation function and the resilient backpropagation algorithm with weight backtracking, with a threshold of 0.01, and up to four of hidden nodes networks were ran 100 times each.

The best neural network, in terms of MSE, for ANN for the returns of type A and B bottles are depicted in Figure 17. In these figures error stands for sum of squared errors, $SSE = n \times MSE$. The results of the best of 100 runs of the ANN are depicted in Table 15. For type A bottles, with three input nodes, the best

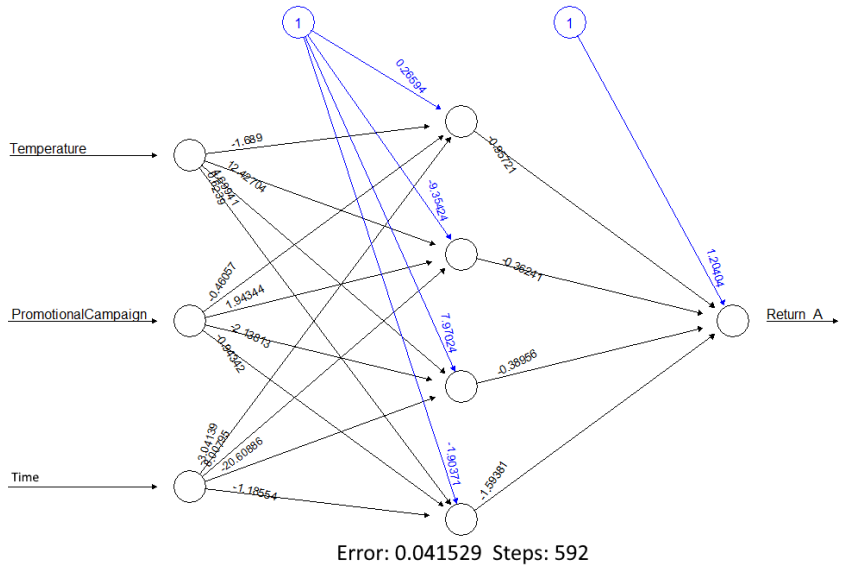

Fig. 17 Best ANN obtained for number of type A

Table 15 Summary of the ANN obtained for type A and B returned bottles.

	Type A	Type B
# nodes	4	4
# steps	592	591
AIC	42.083	34.182
BIC	66.822	54.209
MSE	341 601 46.88	13 52 486.904

model has four hidden nodes and a MSE of 341 601 46.88, 5% of the MSE obtained using TS and 19% of the MSE obtained using MLR. Also four hidden nodes were used in the best ANN model for type B bottles return. The MSE obtained is this model using ANN is also considerably smaller than the one obtained using TS and MLR techniques (46% and 21%, respectively).

The forecast, for the period from Jan/Y2 to Aug/Y2, of the returns of type A and type B bottles were obtained using the best model (see Table 15 and Figure 18). The forecast's behavior in the number of type A and type B bottles resemble the behavior described by the returned number of bottles observed in previous periods and exhibits the increasing tendency, due to the growth of the company in the LPG business.

6.4 Combining the forecasts

Similar to the sales forecast ensemble strategy presented before, for each month, m, it is assumed that the forecast of each algorithm (TS, MLR and ANN) follows a normal probability distribution with mean equal to the punctual value of the

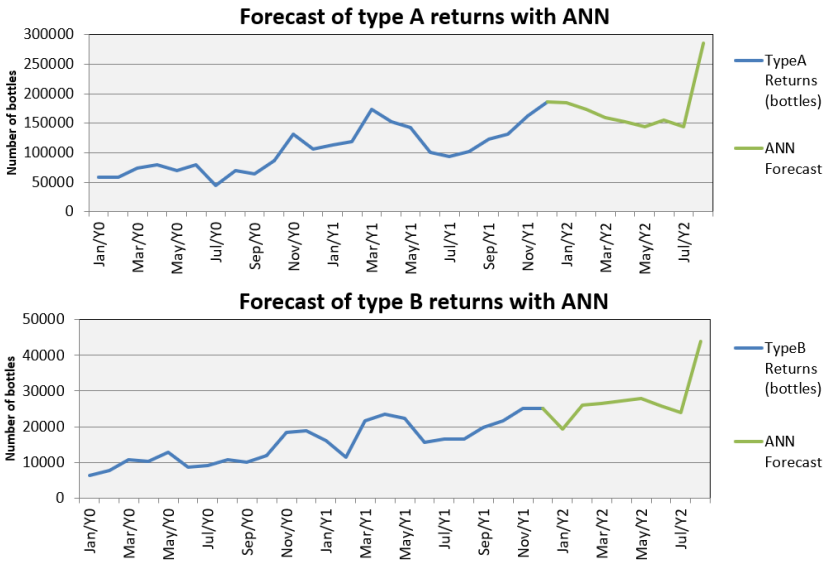

Fig. 18 Forecasts obtained for returned type A and B ALPHA assets (from Jan/Y2 to Aug/Y2).

forecast, $\bar{Y}_{TS}(m)$, $\bar{Y}_{MLR}(m)$, $\bar{Y}_{ANN}(m)$, and variance equal to the MSE, MSE_{TS}, MSE_{MLR}, MSE_{ANN}.

Table 16 Weight assigned to each forecasting method for type A and B number of returned bottles.

Return	α_{TS}	α_{LM}	α_{ANN}
Type A bottles	3.74%	15.52%	80.74%
Type B bottles	27.63%	12.55%	59.82%

The percentual contribution of each forecasting technique for computing the combined forecasts is depicted in Table 16. The standardized weights of each technique are inversely proportional to the magnitude of the forecast errors. For forecasting the returned type A bottles the ANN is the technique that contributes the most, 80.74%, while MLR contributes with 15.52% and the reminder by TS. On the other hand, ANN contributes 59.82% to the forecast of returned type B bottles, while TS contributes with 27.63% and finally MLR with 12.55%. This guaranties that the forecasting techniques with the minor MSE have a larger contribution to the combined forecast, and the techniques with the higher MSE have a minor contribution to the forecast.

Using five Monte Carlo simulation runs and a sample size of 150, it were obtained the forecasts for the months from Jan/Y2 to Aug/Y2, presented in Table 17 and Figure 20.

Figure 19 compares the PDF functions of the forecast for January Year2 of the returns of bottles of type B using the several methods. In blue, the normal distribution that comes from the time series approach, using seasonal indexes

Table 17 ALPHA Returned type A and type B bottles forecast

	Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval Lower bound	Upper bound
ReturnA	Jan/Y2	185 834.60	273.24	185 595.10	186 074.10
ReturnA	Feb/Y2	176 020.50	188.86	175 854.90	176 186.00
ReturnA	Mar/Y2	168 552.40	214.00	168 364.80	168 739.90
ReturnA	Apr/Y2	162 012.40	326.49	161 726.20	162 298.60
ReturnA	May/Y2	153 067.10	284.30	152 817.90	153 316.30
ReturnA	Jun/Y2	163 289.60	778.70	162 607.10	163 972.20
ReturnA	Jul/Y2	146 209.80	270.16	145 973.00	146 446.60
ReturnA	Aug/Y2	263 047.40	865.84	262 288.50	263 806.40

	Month	Average Response of Y	Standard Deviation of Y	95% Confidence Interval Lower bound	Upper bound
ReturnB	Jan/Y2	21 615.39	55.77	21 566.50	21 664.27
ReturnB	Feb/Y2	23 714.37	43.07	23 676.61	23 752.12
ReturnB	Mar/Y2	28 153.74	63.06	28 098.47	28 209.01
ReturnB	Apr/Y2	29 166.76	39.66	29 131.99	29 201.53
ReturnB	May/Y2	29 053.02	54.08	29 005.61	29 100.42
ReturnB	Jun/Y2	24 678.64	72.31	24 615.26	24 742.02
ReturnB	Jul/Y2	23 858.53	131.71	23 743.09	23 973.98
ReturnB	Aug/Y2	36 582.22	71.11	36 519.88	36 644.55

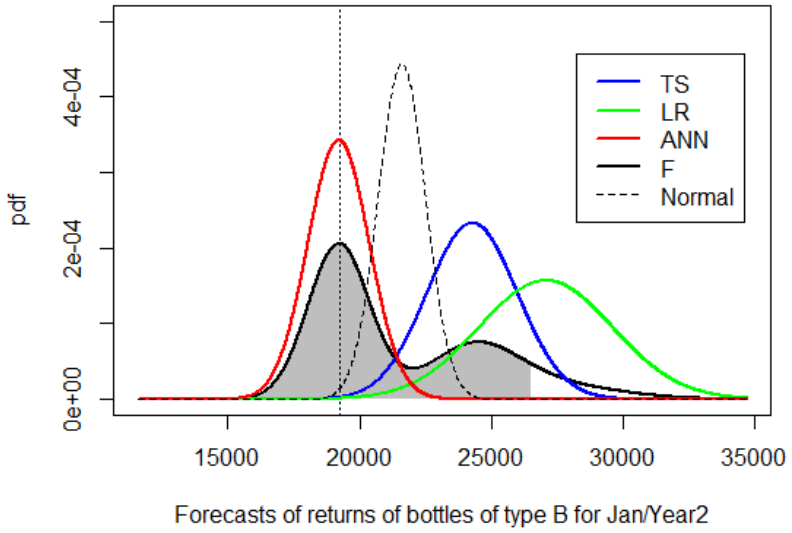

Fig. 19 Curves of PDF functions of the Forecast of type B bottles returns

and Holt's method. In green, the normal distribution obtained from the forecast using the linear regression model for the considered scenario. In red the normal distribution of the forecast using artificial neural networks. The combined function of the three methods according to equation (1) is presented in black. This is the weighted sum of three normal distributions, which is a bit different from a normal distribution obtained with a weighted mean and variance (depicted in dashed

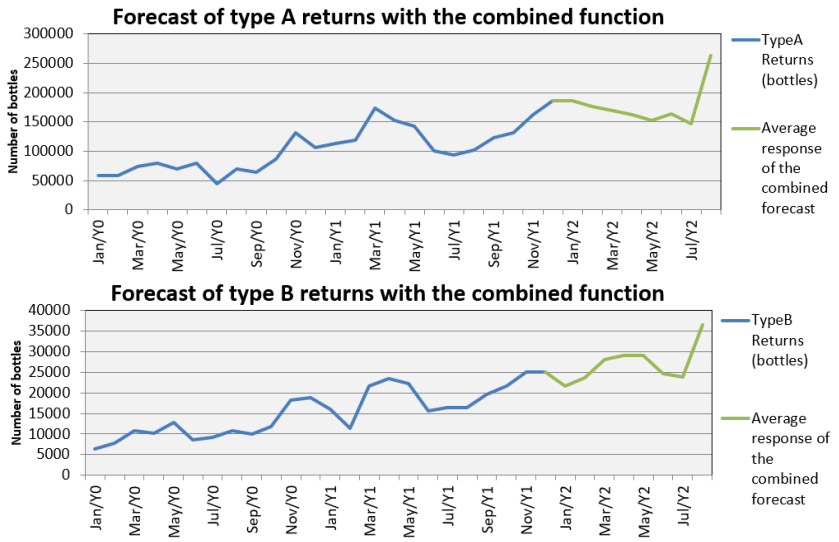

Fig. 20 Combined forecasts for the number of returned type A and B bottles.

black). The 90% confidence interval shaded in grey represents the area under the curve of the combined F.

7 Conclusions and recommendations

The goal of this work was to develop a model for the LPG assets acquisition planning, to answer the industrial challenge proposed by a company in the LPG cylinder business.

For that purpose, it is necessary to forecast the sales of propane gas cylinders, and use it to plan the assets acquisition necessity.

First, time series techniques, namely exponential smoothing and moving averages, were used to compute the seasonal coefficients and to forecast sales and the number of returned bottles.

This approach allowed to see that the national seasonal coefficients are quite distinct from the ones observed for the sales of the company. ALPHA's sales present a larger variability in the seasonal coefficients than the total national sales. For the company the higher coefficients were observed in May, November and December, while the smaller were in July.

Since national and ALPHA seasonality coefficients are different, some other possible explanatory variables should be considered in order to forecast sales with better accuracy. For that reason, several data has been collected, such as atmospheric temperatures, demand in previous periods, objectives of sales, expectation of price increase, and among others variables. Taking into account this data multiple regression models were estimated for the total sales of propane, for the number of type A and B bottles at a national level.

ALPHA wants to find a model to forecast the demand of each type of cylinder. These forecasts are crucial for the company to accurately define an assets acqui-

sition plan, i.e., to determine the amount of LPG cylinders to acquire, and when to acquire them.

Artificial neural networks were used to forecast total propane gas sales and return rates of cylinders (empty bottles).

To conclude, to obtain an improved estimate, these methods were combined in an ensemble method. For this, a probability density function was defined for each method and a Monte Carlo simulation was used. The values obtained are considered in a linear combination, with weights proportional to the accuracy of the method.

The combined method eliminates the drawbacks of individual methods, such as overfitting, and maintains their advantages, leading to more robust forecasts. Furthermore, it allows to deal with non-linearity and seasonality [39].

The company's sales show significant variations, with the seasonal coefficients of the total sales vary from 59% in July to 172% in December, while for type A cylinders from 55% in July until 163% in December, and finally, type B cylinder sales vary from 65% in July to 183% in December. February, May, November and December are the months with high coefficients (above 100%) while June, July, August and September, which correspond to summer season, present the smallest coefficients (below 100%). Furthermore, it was found that the company is increasing the market share, and sales are growing at a rate of 56.8 tons per month.

The MLR suggest that propane national sales are generally decreasing at a rate of 406 tons per month. Futhermore, for each Celsius degree increase in the temperature, the national sales would decrease nearly 562 tons. Relatively to the company's sales, it was observed that for the months when the company's distributors have to meet sales objectives there is a significant impact on sales of almost 2 thousand tons. Also, for the months where there is an expectation of price increase, sales increase 1458 tons. When considering all other variables constant, for each Celsius degree increased, the company's sales decrease 51 tons, while for each degree that the temperature drops, sales will increase 51 tons.

The best neural network, for the company total sales presents three hidden nodes and a MSE of 34024.6106, which is smaller than the one obtained using TS and MLR models. For type A and type B bottles sales, the ANN have five and four input nodes, respectively, three hidden nodes and has a MSE considerably smaller than the ones obtained using TS and MLR methods. In fact, the MSE obtained using ANN is around 30% of the MSE obtained using MLR (both for type A and type B bottles sales), and is 10% of the MSE obtained using TS for type A bottles and 20% for type B bottles.

The forecasts obtained using the different individual methods presented lower values than the ones estimated by the company. Also, for the combined forecast method, the average forecast is also below the company expectations. In fact, the company sales predictions are nor included in the 95% confidence intervals found using the proposed methodology. This means that the company's expectation of growth is larger than the ones predicted the model here proposed which may indicate that the company should reflect on their strategy. However, since the authors do not have information concerning the company's growth strategy, perhaps the company is aiming at further promotional campaigns or other actions that are expected to increase the volume of sales, that have not been considered in the methodology used in the present work.

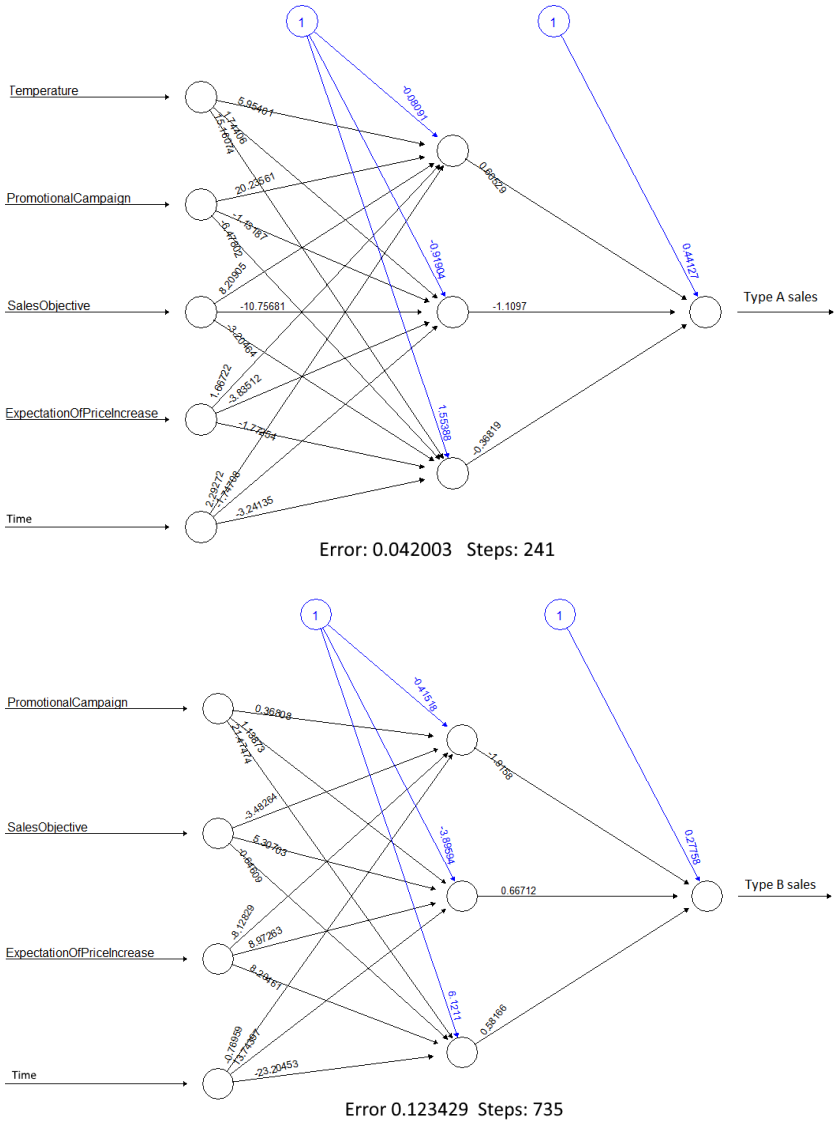

Fig. 21 The best ANN for propane type A and B bottles.

Acknowledgements [Removed for review]

Conflict of interest

The authors declare that they have no conflict of interest.

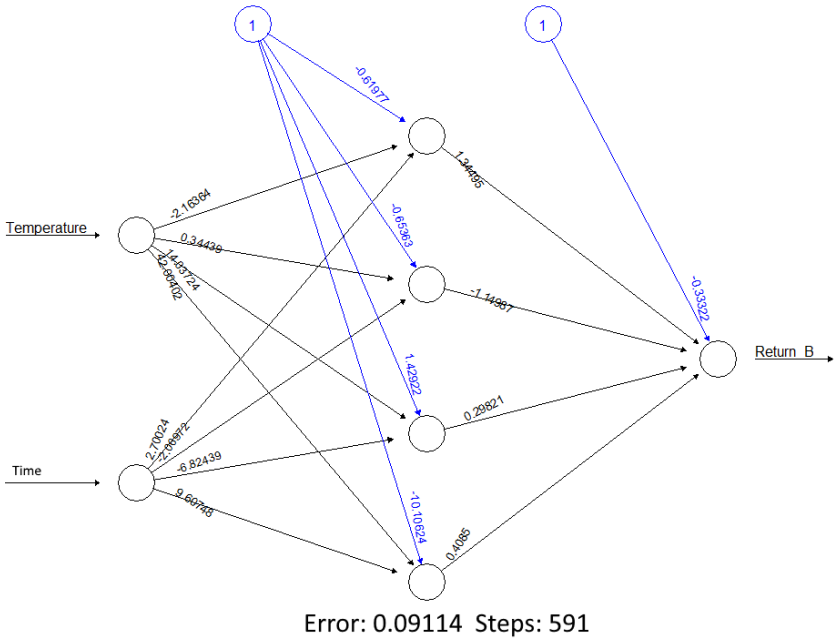

Fig. 22 Best ANN obtained for number of type B returned bottles

References

1. Adhikari, R., Agrawal, R.K.: A linear hybrid methodology for improving accuracy of time series forecasting. Neural Computing and Applications **25**(2), 269–281 (2014)
2. Amasyali, K., El-Gohary, N.M.: A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews **81**(Part 1), 1192–1205 (2018)
3. Aras, H., Aras, N.: Forecasting residential natural gas demand. Energy Sources **26**(5), 463–472 (2004)
4. Balestra, P., Nerlove, M.: Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas. Econometrica **34**(3), 585–612 (1966)
5. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer Science & Business Media (1991)
6. Burnham, K.P., Anderson, D.R.: Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media (2003)
7. Carrasco-Gallego, R., Ponce-Cueto, E.: Forecasting the returns in reusable containers' closed-loop supply chains. a case in the lpg industry. In: 3rd International Conference on Industrial Engineering and Industrial Management XIII Congreso de Ingeniería de Organización, pp. 311–320. Universitat Politécnica de Catalunya (2009)
8. Cassettari, L., Bendato, I., Mosca, M., Mosca, R.: A new stochastic multi source approach to improve the accuracy of the sales forecasts. foresight **19**(1), 48–64 (2017)
9. Dombayci, O.A.: The prediction of heating energy consumption in a model house by using artificial neural networks in denizli-turkey. Advances in Engineering Software **41**(2), 141–147 (2010)
10. Draper, N., Smith, H.: Applied Regression Analysis. Wiley (1998)
11. Erdogdu, E.: Natural gas demand in turkey. Applied Energy **87**, 211–219 (2010)
12. Feng, C., Cui, M., Hodge, B.M., Zhang, J.: A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. Applied Energy **190**, 1245–1257 (2017)

13. Fernández, J.C., Cruz-Ramírez, M., Hervás-Martínez, C.: Sensitivity versus accuracy in ensemble models of artificial neural networks from multi-objective evolutionary algorithms. Neural Computing and Applications **30**(1), 289–305 (2018)
14. Fonseca, S.: Characterization of the energy consumption in portugal's residential sector. Master's thesis, Instituto Superior Técnico, Lisbon, Portugal (2014)
15. Freedman, D.A.: Statistical Models: Theory and Practice. Cambridge University Press (2009)
16. Gardner, E.S.: Exponential smoothing: The state of the art. Journal of forecasting **4**(1), 1–28 (1985)
17. Hamilton, J.D.: Time Series Analysis. Princeton University Press (1994)
18. Holt, C.C.: Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting **20**(1), 5–10 (2004)
19. Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D.: Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media (2008)
20. Hyndman R., A.G.: Forecasting: Principles and practice. `http://otexts.com/fpp/` (2012). Accessed: 2017-09-27
21. Jiang, Y., Yin, S., Kaynak, O.: Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond. IEEE Access **6**, 47,374–47,384 (2018)
22. Lantz, B.: Machine learning with R. Packt Publishing Ltd (2013)
23. Liu, L.M., Lin, M.W.: Forecasting residential consumption of natural gas using monthly and quarterly time series. International Journal of Forecasting **7**, 3–16 (1991)
24. Mentzer, J.T., Cox, J.E.: Familiarity, application, and performance of sales forecasting techniques. Journal of Forecasting **3**(1), 27–36 (1984)
25. Montgomery, D.C., Johnson, L.A., Gardiner, J.S.: Introduction to Linear Regression Analysis, 2nd edn. McGraw-Hill (1990)
26. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, 5th edn. Wiley (2012)
27. Montgomery, D.C., Runger, G.C.: Applied statistics and probability for engineers. John Wiley & Sons (2010)
28. Naderpour, H., Mirrashid, M.: Shear failure capacity prediction of concrete beam–column joints in terms of anfis and gmdh. Practice Periodical on Structural Design and Construction **24**(2), 04019,006 (2019)
29. Naderpour, H., Mirrashid, M., Nagai, K.: An innovative approach for bond strength modeling in frp strip-to-concrete joints using adaptive neuro–fuzzy inference system. Engineering with Computers pp. 1–18 (2019)
30. Sánchez-Úbeda, E., Berzosa, A.: Modeling and forecasting industrial end-use natural gas consumption. Energy Economics **29**(4), 710–742 (2007)
31. Soldo, B.: Forecasting natural gas consumption. Applied Energy **92**, 26–37 (2012)
32. Sousa, J.: Background of portuguese domestic energy consumption at european level. In: IT4Energy International Workshop on Information Technology for Energy Applications (2012)
33. Thaler, M., Grabec, I., Poredoš, A.: Prediction of energy consumption and risk of excess demand in a distribution system. Physica A: Statistical Mechanics and its Applications **355**(1), 46–53 (2005)
34. Tonković, Z., Zekić-Sušac, M., Somolanji, M.: Predicting natural gas consumption by neural networks. Tehnicki Vjesnik **16**(3), 51–61 (2009)
35. Vitullo, S.: Disaggregating time series data for energy consumption by aggregate and individual customer. ProQuest (2011)
36. Vitullo, S.R., Brown, R.H., Corliss, G.F., Marx, B.M.: Mathematical models for natural gas forecasting. Canadian applied mathematics quarterly **17**(7), 807–827 (2009)
37. Vondráček, J., Pelikán, E., Konár, O., Čermáková, J., Eben, K., Malý, M., Brabec, M.: A statistical model for the estimation of natural gas consumption. Applied Energy **85**(5), 362–370 (2008)
38. Wright, D.J.: Forecasting data published at irregular time intervals using an extension of holt's method. Management Science **32**(4), 499–510 (1986)
39. Yang, Y., Chen, Y., Wang, Y., Li, C., Li, L.: Modelling a combined method based on anfis and neural network improved by de algorithm: A case study for short-term electricity demand forecasting. Applied Soft Computing **49**, 663–675 (2016)
40. Yin, S., Jiang, Y., Tian, Y., Kaynak, O.: A data-driven fuzzy information granulation approach for freight volume forecasting. IEEE Transactions on Industrial Electronics **64**, 1447–1456 (2017)