FUEL RETAIL MARKET: ASSESSING THE DETERMINANTS THAT INFLUENCES THE PERFORMANCE OF SALES OF FUEL STATIONS

Stephanie Espadilha
ESTGA - Águeda School of Technology and Management,
University of Aveiro, Portugal
sespadilha@ua.pt

Marco Costa
ESTGA - Águeda School of Technology and Management,
CIDMA - Center for Research & Development in Mathematics and Applications,
University of Aveiro, Portugal
marco@ua.pt

Daniel Magueta
ESTGA - Águeda School of Technology and Management,
GOVCOPP - Research Unit on Governance, Competitiveness and Public Policies,
University of Aveiro, Portugal
dmagueta@ua.pt

ABSTRACT

The oil and energy sector is a very traditional, controversial and competitive sector. This study is based on a Portuguese fuel company and its main objective is to identify and characterize potential variables with predictive capacity for sales of new fuel stations. The database consists of a set of context variables with predictive potential for sales of fuel stations and monthly sales in terms of fuel volume. The research methodology focused on statistical methods of exploratory data analysis, clusters analysis and regression models. The fuel station context variables tend to characterize the socio-economic conditions of the area of influence of each station, such as population density variable, others related to the similar existing supply of both the company itself and the competing companies, and others related to geographical location and accessibility. The exploratory data analysis allowed to identify several patterns in the time series of sales indicating that the investigation of factors must be segmented. Homogeneous groups of fuel stations were identified through a hierarchical agglomerative clustering procedure considering the Ward's minimum variance method and the square Euclidian distance as distance measure. For each of the groups identified, multiple linear regression models were adjusted considering the annual fuel sales in the 1st, 2nd and 3rd years of operation of the stations as dependent variables. The results show that not all the exogenous variables are statistically significant. However, it is possible to conclude that the average daily traffic is the variable with predicted capacity for the most of the groups of fuel stations analyzed.

Keywords: *Fuel Retail Market, Fuel stations, Multivariate analyses, Sales determinants, Sales forecasting*

1. INTRODUCTION

To model time series sales behavior with the intent to be a useful tool to forecasting is a great challenge. All types of energy consumption, including car fuel, are growing in almost every country in the world, forced by several factors such as the increase in population and the search for better living conditions. In this context, the capacity to model and forecast with adequate techniques is essential for accurate planning, production and distribution capacities by the firm's operation in these areas. The development of trustworthy models is difficult for a variety of reasons, but usually, one of the biggest problems are related to the difficulty of having access to necessary information and data. Nowadays, markets in almost every activity sectors are increasingly competitive, and the fuel sector is no exception, particularly at the retail moment. In this way, it is fundamental for companies competing in this area, to adopt management strategies that could allow them to enhance the competitive advantage they may have against the competition. The ability to take advantage of all the information they generate, transforming raw data into useful information to help management level is essential for companies to succeed and face the enormous challenges ahead. The understanding of the phenomenon that guide and influence sales are a key element in the decision-making process of management activities. This can make firms to more accurately make sales forecasting and in that way also predict the major factors that influence their performance and results. In this work, the main objective is to identify and characterize potential variables with the capacity to explain the sales behavior of the fuel stations in Portugal. The work will have the following structure. In section 2 it is made a brief literature review about some of the models used. Section 3 shows the methodology that was used throughout this study and explains the data collection process. Section 4 analyses the results. In this part, we used multiple linear regression models and variable selection techniques for three different groups, in the first, second and third year of activity sales, in order to find variables with predictive capacity. Finally, Section 5 refers to some conclusions about the results.

2. LITERATURE REVIEW

Forecasting plays an essential role in the planning process of all future activity, decision making and control in any company. The uncertainty related to many temporal phenomena does not allow the exact knowledge of its behavior in the future, and therefore, leads us to the necessity to make predictions (Caiado, 2016). Being able to accurately predict sales allow firms to improve market performance, minimize profit losses and to plan manufacturing processes and marketing policies more efficiently (Fantazzinia and Toktamysova, 2015). This is particularly relevant in this market, were the uncertainty and the risk companies face are phenomena at a global scale. In order to mitigate the problems that could derive from there, and do a more careful and efficient supply chain and operation management, to accurately be able to forecast sales is crucial in fuel retail activity. At a more organizational level, sales forecasts are necessary as a crucial input to decision activities in many functional areas, for instance in marketing, sales, production, purchasing, finance and accounting (Mentzer and Bienstock, 1998). An accurate sales model permits more efficient inventory management and has long been recognized that provide the basis for firms' delivery and refill procedures. The connection among retail stocks and sales is examined at an aggregate level and has been found that effective inventory management is contingent to a large extent on the correct estimating of retail sales Barksdale and Hilliard (1975). This is also what Thall (1992) and Agrawal and Schorling (1996) explain about the importance of demand predicting, since they consider that plays a serious role in the profit of retail operations and point out that a poor forecast would result in too much or too little stocks directly affecting firm revenue and the competitive position of the retail business. Montgomery et al. (2008), point out that the main objective of using forecasting methods is to predict future events, with the aim of reducing risk in decision making. They also argue that the greater availability of resources in the implemented forecasting method could allow being able to improve the accuracy of the forecast and thus reduce some of the losses resulting from uncertainty in the decision-making process. Even though Gas stations fuel sales is not a very addressed topic in the literature, there are many studies trying to model the behavior in other related markets, namely in the energy sector, are much more common and share some of the specifics. Harris and Lon-Mu (1993) have analyzed the relationship between electricity consumption using a number of potentially relevant variables, such as weather, price and consumer income, finding out a high seasonality of electricity demand.

Analyzing the consumption pattern of electrical energy, Ranjan and Jain (1999) apply multiple linear regression models and population and weather parameters as explanatory variables of the energy consumption for different seasons. Bianco et al. (2013) studied the residential and non-residential annual electricity consumption applying simple and multiple regression models using historical electricity consumption, gross domestic product (GDP), GDP per capita and population has independent variables. Their results demonstrate that the selected explanatory variables are strongly correlated to the electricity consumption, and in some of the models the coefficient of determination for the simple and multiple regression models was extremely high (0.975 and 0.990). Using stepwise selection techniques, like the ones applied in this paper, Filippína et al. (2013) applied multivariate analysis to assess the historical consumption of natural gas for heating in multifamily buildings. With data from 72 apartments from different buildings, with different orientations, different energy groups (clusters) were generated. After the stepwise method applied the authors to select the variables categorizing the annual energy consumption. In a study made about the selection of location for gas stations Semih and Seyhan (2011) points out that this is one of the key factors of any business success, and for that reason also in the fuel sector. According to them, the problem of fuel station selection involves numerous quantitative and qualitative factors, such as the number of other stations in the area, traffic directions, social composition of surrounding residential area, and curb appeal of the station structure.

3. METHODOLOGY AND DATA

The data used in the paper is originated from a set of variables with theoretically explanatory potential to model fuel sales, namely gasoline and diesel, at a retail level. The information is related to forty-eight fuel stations, located in Portugal mainland. The dependent variable, whose behavior is sought to be explained, is the monthly sales in terms of fuel volume, measured in cubic meters. The independent variables associated with each fuel station tend to characterize the socioeconomic conditions of the area of influence of each location, such as population density and purchasing power, competition issues, namely the existence and proximity of a rival firm and rivals' company's geographical location and accessibility. Regarding the population density, this is represented in two variables, population density by town and population density by county, in both cases, this is measured by the ratio between the number of inhabitants and the area of the zone. The purchasing power in the ratio of county purchasing power to the national average. In relation to the distance to the nearest arterial route, that is, the distance to the nearest motorway, it is visible in two variables, linear distance, and real distance, both measured in kilometers. The competition is also presented by two variables, the number of the low-cost gas station within a 10 km radius and the number of regular gas stations within a 2 km radius. In table 1 are shown the description of the variables and the basic summary descriptive statistics.

Table following on the next page

Table 1: Variables description and summary descriptive statistics

VARIABLES	Mean	Median	Standard deviation
Average daily traffic (no. vehicles per day)	8703.19	8042.00	5105.609
Population density of the town (hab/km^2)	2230.95	481.47	3896.481
Population density of the municipality (hab/km^2)	1296.20	399.95	1877.244
Ratio between the purchasing power of the municipality and the national average (%)	98.8850	93.7400	26.20711
Linear distance to nearest arterial pathway (km)	2.117	2.000	1.5904
Actual distance to nearest arterial line (km)	3.252	3.000	2.2317
No. of low cost, within a 10 km radius	2.90	3.00	1.708
No. of gas station, within a 2 km radius	2.71	2.00	3.445
Average monthly sales in the 1st year of activity (m^3)	213.4774	168.2260	152.71901
Average monthly sales in the 2nd year of activity (m^3)	267.9661	226.9925	166.69976
Average monthly sales in the 3rd year of activity (m^3)	287.8882	247.0095	173.11307

The methodology used is focused on statistical methods of exploratory data analysis, using descriptive statistics, cluster analysis and regression models, in particular multiple linear regression models. The variables selection techniques used were stepwise, forward and backwise. The exploratory analysis of the data allowed to identify several patterns in the time series of sales, indicating that the investigation of factors must be segmented. Through the SPSS software, homogeneous groups of fuel stations were identified through a Agglomerative Hierarchical Clustering (AHC) procedure considering Ward's minimum variance method and the square euclidean distance as a measure of distance. Subsequently, the monthly sales of the respective gas stations were grouped into these four groups that arose through the clustering procedure. In order to facilitate the analysis and reading of the data, for each group the average monthly sales of the gas stations belonging to each group were calculated. Through the analysis of figure 1, it is possible to verify that the gas stations were grouped by the sales volume. Group 1 and group 2 initially have very close sales, however, as of month 36, the sales of group 2 increase significantly. Comparing the last month, group 2 sells 65% more than group 1. Group 3 sells 177% more than group 1 and 68% more than group 2. Finally, group 4 that is only constituted by gas station 9 is the one that sells the most, not been grouped because its sales are much higher than the rest.

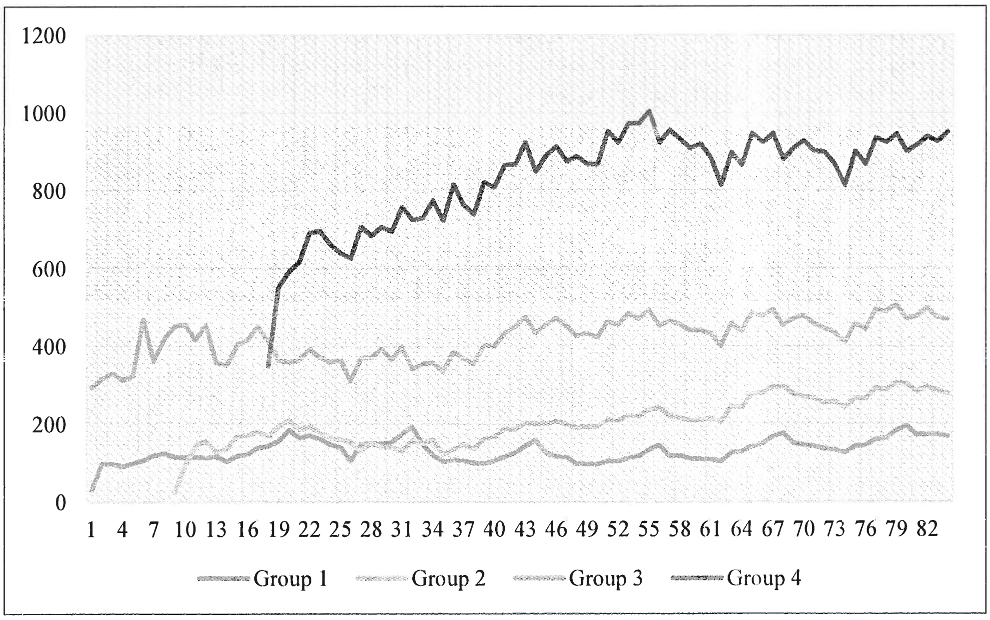

Figure 8: Sales by four group (gas stations average within each group)

For each group, table 2 was constructed with the mean and median of each variable. From this table we can analyze that the value of the variables of the average daily traffic, the population density per town and the population density per municipality increases throughout the three groups, showing a positive relation with the sales volume, which is in line with intuition regarding the importance of the number of potential consumers as an explanation of sales behavior. The number of low costs within a 10 km radius and the number of non-discount competing gas stations within a 2 km radius is also increasing throughout the three groups, although this behavior may seem anti-intuitive, i.e. more competition and more sales, this proves that the weight the number of consumers has in this business area is extremely important.

Table 2: Variables by group

Variables	Group 1		Group 2		Group 3	
	Mean	Median	Mean	Median	Mean	Median
Average daily traffic (ADT)	6336.14	6197.5	8420.93	7232	11611.46	11225
Population density of the town (PDT)	675.92	142.18	1176.11	485.14	4073.68	1940.28
Population density of the municipality (PDM)	371.37	66.3	553.43	399.95	2609.82	2173.6
Ratio between the purchasing power of the municipality and the national average (PP)	86.3971	87.215	86.2766	91.835	118.997	101.36
Linear distance to nearest arterial pathway (LD)	2.179	1.95	2.193	2.05	2.1	2
Actual distance to nearest arterial line (ALD)	3.529	3.1	3.329	3.25	3.069	2.5
No. of low cost, within a 10 km radius (N10)	2.43	2	2.86	3.5	3.46	4
No. of gas station, within a 2 km radius (N2)	0.93	0	2.5	2	4.08	3
Average monthly sales in the 1st year of activity (AVG1)	90.74	92.721	173.4043	168.6855	357.0362	338.445
Average monthly sales in the 2nd year of activity (AVG2)	120.7094	125.0495	225.7085	226.9925	433.6079	414.946
Average monthly sales in the 3rd year of activity (AVG3)	134.6162	140.0925	245.7799	247.0095	451.2858	406.169

On the other hand, the value of the purchasing power variables, the linear distance to the nearest arterial pathway and the actual distance to the nearest arterial pathway, do not show differences between groups, but may also contain predictive capacities.

4. RESULTS AND ANALYSES

Population density per town and population density per municipality are strongly correlated, so the introduction of both variables in the model would generate multicolinearity problems. The same happens with the linear distance to the nearest arterial pathway and the actual distance to the nearest arterial pathway. The choice of variables to be placed in the model as potential explanatory variables was made using the ordinal correlation of these variables with the independent variable (sales). As can be seen in table 3, for group 1, the variable population density of the town has higher correlations than the population density of the municipality. Relatively to distances to the nearest arterial pathway, the actual distance to the nearest arterial pathway has correlations higher than the linear distance to the nearest arterial pathway. Compared to the population density of group 2, the population density variable of the parish also has correlations higher than the population density of the county, in the monthly sales of the 1st and 2nd year of activity.

Table 3: Spearman's ordinal correlation coefficient by group

	ADT	PDT	PDM	PP	LD	ALD	N10	N2
Group 1								
AVG1	.436	.602	.411	.266	.381	.456	.274	.019
	(.081)	(.010)	(.101)	(.302)	(.179)	(.101)	(.287)	(.943)
AVG2	.412	.410	.199	.307	.355	.443	.179	.007
	(.100)	(.102)	(.444)	(.231)	(.213)	(.113)	(.492)	(.980)
AVG3	.216	.321	.228	.363	.145	.242	.333	.145
	(.405)	(.208)	(.378)	(.152)	(.620)	(.404)	(.192)	(.578)
Group 2								
AVG1	-.241	.054	-.088	-.572	.053	-.119	.103	-.182
	(.352)	(.837)	(.736)	(.016)	(.856)	(.684)	(.695)	(.484)
AVG2	-.030	.126	.063	-.480	.102	.110	.336	-.237
	(.911)	(.629)	(.811)	(.051)	(.728)	(.707)	(.187)	(.360)
AVG3	.055	-.285	-.246	-.219	-.193	-.217	.221	.076
	(.833)	(.268)	(.342)	(.399)	(.508)	(.457)	(.395)	(.772)
Group 3								
AVG1	.696	.553	.574	.061	.345	.328	-.242	.272
	(.008)	(.050)	(.040)	(.844)	(.248)	(.274)	(.425)	(.369)
AVG2	.715	.539	.577	.074	.420	.413	-.339	.227
	(.006)	(.057)	(.039)	(.809)	(.153)	(.160)	(.257)	(.455)
AVG3	.583	.622	.585	.072	.199	.187	-.114	.386
	(.036)	(.023)	(.036)	(.816)	(.515)	(.540)	(.711)	(.193)

The values in parentheses are the p-value.

For group 3, it is the population density variable of the municipality that has higher correlations to the monthly sales of the 1st and 2nd year of activity, however, in the monthly sales of the 3rd year, is the variable population density of the town that has higher correlations. Regarding the distances to the nearest arterial pathway, in this group is the linear distance to the nearest arterial pathway that has higher correlations than the actual distance to the nearest arterial pathway. For each of the identified groups, the models were adjusted considering the fuels annual sales in the 1st, 2nd and 3rd years as dependent variables. From the modeling point of view, statistically significant variables were those that have coefficients associated with p-values lower than 0.2 since this is a study with socio-economic variables and regional/local context. Analyzing table 4, for the first year of group 1, for each kilometer that the actual distance to the nearest arterial route increases, sales in the first year of activity increase by 11.106 m^3 and by each number of cars that average daily traffic increases, sales in the first year of activity increase by 0.006 m^3. For the second year of group 1, for each kilometer that the actual distance to the nearest arterial route increases, sales in the second year of activity increase by 8.417 m^3 and by each number of cars that average daily traffic increases, sales in the second year of activity increase by 0.009 m^3. Relatively to the sales of the third year, analyzing table 4 we can conclude that of the variables studied, there is none that has explanatory capacity.

Table following on the next page

Table 4: Coefficients of group 1

Sales	1º year	2º year	3º year
(Constant)	11.905 (0.711)	36.180 (0.286)	–
ALD	11.106 (0.030)	8.417 (0.095)	–
ADT	0.006 (0.143)	0.009 (0.058)	–
R^2	**0.442**	**0.416**	–
DW	**1.960**	**1.854**	–

The values in parentheses are the p-value.

The coefficient of determination for the first year of sales is $R^2 = 0.442$, that is, 44.2% of the variance of sales in the first year is explained by the model. For the second year of sales, since $R^2 = 0.416$, this means that 41.6% of the sales variance in the second year is explained by the variables actual distance to the nearest arterial route and average daily traffic. The Durbin-Watson test for the first and the second year of sales are close to two, so it's possible to accept that the residuals are independent, that is, there is no autocorrelation issues that may be concerning. Looking at table 5, for the first year of group 2, if the ratio of the purchasing power of the municipality increases by 1pp, sales in the first year of activity decreased by 3.185 m^3, for each increase of one new low cost competitor within a 10 km radius, sales in the first year of activity increase by 11.321 m^3 and for each unit increase in the population density of the town, sales are expected to increase by 0.011 m^3. For the second year, for each perceptual point increase in the purchasing power ratio, sales decreased by 4.484 m^3 and the increase of a low cost competing gas station in a 10 km radius will result in an increase in sales of 24.678 m^3. Relative to the sales of the third year, none of the studied variables has predictive capacity.

Table 5: Coefficients of group 2

Sales	1º year	2º year	3º year
(Constant)	430.037 (0.000)	580.924 (0.001)	–
PP	-3.185 (0.004)	-4.484 (0.013)	–
N10	11.321 (0.040)	24.678 (0.020)	
PDT	0.011 (0.48)	–	–
R^2	**0.585**	**0.496**	-
DW	**1.243**	**1.666**	-

The values in parentheses are the p-value.

Both the first year and the second year, the variables ratio between the purchasing power of the county and the national average and the number of low costs are statistically significant to model. The coefficient of determination for the first year is $R^2 = 0.585$, that is, 58.5% of the variance of sales in the first year is explained by the variables of purchasing power ratio, competitor within a 10 km radius and town density. For the second year of sales $R^2 = 0.496$, this means that 49.6% of the variance of sales in the second year is explained by the independent variables, purchasing power ratio and number of competing low costs. Unlike the other two groups, the variable selection techniques used in group 3 do not achieve the same results, which may mean some lack of robustness since they depend on the technique used. In this way, the results of group 1 and group 2 turn out to be more robust because regardless of the technique used the results are always the same.

The results for group 3, using the stepwise forward regression technique, are presented in table 6. In those result is possible to see that in the first year of sales, for each unit increase in the average daily traffic, sales increase by 0.020 m^3, for each increase of inhabitants per km^2, sales increase by 0.054 m^3, for each increase of competing fuel gas station within a 2km radius, sales decrease by 14.142 m^3 and for every kilometer that the linear distance to the nearest arterial route increases, sales decrease by 17.845 m^3. For the second year, for each increase in average daily traffic, will result in an increase of sales by 0.013 m^3, for each increase of inhabitants per km^2, sales increase by 0.041 m^3 and per every increase of a competing fuel station within a 2km radius, sales decrease by 9.353 m^3. For the third year, for each car per day increases, sales will increase by 0.009 m^3, for each unitary increase of inhabitants per km^2, sales increase 0.024 m^3.

Table 6: Coefficients of group 3 – Stepwise Forward

Sales	1º year	2º year	3º year
(Constant)	66.586 (0.261)	200.960 (0.008)	261.791 (0.001)
ADT	0.020 (0.001)	0.013 (0.007)	0.009 (0.027)
PDM	0.054 (0.001)	0.041 (0.008)	0.024 (0.036)
N2	-14.142 (0.039)	-9.353 (0.161)	–
LD	-17.845 (0.194)	–	–
R^2	**0.859**	**0.728**	**0.541**
DW	**2.053**	**2.7**	**2.683**

The values in parentheses are the p-value.

The coefficient of determination for the first year of sales is $R^2 = 0.859$, that is, 85.9% of the variance of sales in the first year is explained by the four variables in the model. This explanatory capacity of the model reveals a very good fit. For the second year the $R^2 = 0.728$, this means that 72.8% of the variance of sales is explained by the variables average daily traffic, population density of the municipality and number of competing stations within a 2 km radius. For the third year the $R^2 = 0.541$, that is, 54.1% of the variance of sales is explained by the variables average daily traffic and population density of the municipality.

5. CONCLUSION

The main purpose of this study was to find a set of variables with explanatory capabilities that could model the behavior of the sales made by gas stations. Since gas stations are very heterogeneous it was important to group them into clusters. In this way it's possible to analyze gas stations with similar characteristics, and in that way obtain the best adjustment to the possible model, allowing a better understanding of the mechanism behind the evolution of diesel and gasoline sales. After applying stepwise regression techniques, we conclude that for group 1, for both first and second year sales, variables with predictive capacity are the actual distance to the nearest arterial route and average daily traffic. For group 2 in the first year of sales, the variables with explanatory capacity are the purchasing power ratio of the municipality, the number of competing low cost gas stations within a 10 km radius and the population density of the town. For the second year of sales, only the variable ratio between the purchasing power of the municipality and the number of competing low costs are maintained in the model. For the sales of the first year of group 3, the variables with predictable capacity are the average daily traffic, the population density of the municipality, the number of competing stations and the linear distance to the nearest arterial route, however, for sales in the second year, the variables with predictive capacity are the average daily traffic, the population density of the county and the number of competing stations. Finally, for the third year sales, there are only two variables with predictive capacity, namely average daily traffic and population density of the municipality. Overall, the average daily traffic is the variable which shows explanatory capacity in the majority of the gas station groups and models applied. This is a very interesting result since it would be expected that many other of the available variables could have some importance explaining the behavior of gas station sales. The purchasing power of the inhabitants of a local would be expected to have more relevance, but the variable didn't reveal to be statistical significant in almost every model. This could be in part explained by the use Portuguese make of their cars, mainly for travelling to work, due to the lack of public transportations options. In this context, even in poorer regions, gas stations can achieve similar sales when compared with regions with more buying capacity.

ACKNOWLEDGEMENT: *Marco Costa was partially supported through Portuguese funds by the Center for Research and Development in Mathematics and Applications (CIDMA) and the Portuguese Foundation for Science and Technology ("Fundação para a Ciência e a Tecnologia" [FCT]), within project UID/MAT/04106/2019. Daniel Magueta was partially supported by the research unit on Governance, Competitiveness and Public Policy (UID/CPO/04058/2019), funded by national funds through FCT - Fundação para a Ciência e a Tecnologia.*

LITERATURE:

1. Agrawal, D. & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing,* 72 (4), pages 383–407.
2. Barksdale, H. & Hilliard, J. (1975). A cross-spectral analysis of retail inventories and sales. *Journal of Business.* 48 (3), pages 365–382.
3. Bianco, V., Manca, O. & Nardini, S. (2013). Linear regression models to forecast electricity consumption in Italy. *Energy Sources,* Vol 8, pages 86–93.
4. Caiado, J. (2016). Métodos de Previsão em Gestão – Com Aplicações em Excel (2ª Edição). Lisboa: Edições Sílabo.
5. Fantazzinia, D. & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics,* Vol. 170, Part A, pages 97-135.
6. Filippína, C., Ricard, F. & Flores L. (2013). Evaluation of heating energy consumption patterns in the residential building sector using stepwise selection and multivariate analysis. *Energy Build,* Vol 66, pages 571–81.
7. Harris, J. & Lon-Mu, L. (1993). Dynamic structural analysis and forecasting of residential electricity consumption. *International Journal fo Forecasting,* Vol 9, Issue 4, pages 437-455.
8. Mentzer, J. & Bienstock, C. (1998). Sales Forecasting Management. Sage, Thousand Oaks, CA.
9. Montgomery, D., Johnson, L. & Gardiner, J. (2008). Introduction to Time Series Analysis and Forecasting. John Wiley & Sons. New Jersey.
10. Ranjan M. & Jain V. (1999) Modelling of electrical energy consumption in Delhi. *Energy,* Vol 24(4) pages 351–61.
11. Semih, T. & Seyhan, S. (2011). A Multi-Criteria Factor Evaluation Model for Gas Station Site Selection. *Journal of Global Management, 2 (1),* pages 12-21.
12. Thall, N. (1992). Neural forecasts: A retail sales booster. *Discount Merchandiser,* Vol.23 (10), pages 41–42.