International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume 6, Issue 3: Page No.74-83 ,May-June 2017
http://www.mnkjournals.com/ijlrst.htm

MODEL ESTIMATION AND TREND FOR THE SALE OF TRUCK SPARE PARTS USING LEAST SQUARES AND BAYESIAN NETWORKS

[1]Juan Manuel Flores Santillán, [2]Tonahtiu Arturo Ramírez Romero,[3]Miguel Patiño Ortíz
[1,3]SEPI-ESIME Zacatenco, Instituto Politécnico Nacional, Cd. México, México
[2]Instituto Politécnico Nacional, México

Abstract*–The estimation and trend for the sale of truck spare parts over twelve months horizon is a crucial planning for all those people who have a managerial or strategic position in any company which is dedicated to merchandise this kind of items. A model estimation based on least squares and Bayesian Networks (MBN) is developed using six variables which are identified economic indicators and daily sale of truck spare part variable is calculated in order to get its trend. The empirical results suggest that MBN can significantly improve the estimation for sale of truck spare part for 12-month ahead prediction in order to generate a strategy with others areas such as marketing, purchasing and logistics.

Keywords–Model estimation, sales of spare parts, least squares, Bayesian networks

I. INTRODUCTION

The merchandising of anything, invariably involves establishing a quantitative sales target and intrinsically controlling a purchase objective responsibly, and of course considering other aspects such as logistics, etc., which for the time being will not be thoroughly discussed, but otherwise a marketing plan that does not translate into a strategy and a consistent goal, represented by concrete actions, is not a true plan of work but a catalog of good intentions or desires. This research aims to present a commercial model of support for all those who have a managerial or strategic position where there is a need to have an estimate or trend of sale in the sale of truck spare parts in Mexico, in the scope of heavy-duty dealers, (i.e. trucks or buses).

II. RELATED RESEARCH

The motor vehicle sector in Mexico is an important part of the country economy because it represents 3.5% of Gross Domestic Product and it accounts for 19.8% of manufacturing sector. It has an important presence in whole country due to it has developed big manufacturing clusters on the north and center region and important networking of distribution. It has manufacturer plants in twelve states for assembling light vehicles and engines; it has manufacturer plants in eight states for assembling heavy trucks and engines and it has manufacturer providing enterprises of spare parts in twenty six states [5]. Mexico´s position in the global motor vehicle sector is not negligible; it belongs to NAFTA between United States of America and Canada.

Sale estimation is, in general, an interesting and difficult task. Many dealers or independent distributors struggle reduce their logistic costs and increase profits, and to have an accurate estimation model for the sale of truck spare parts is an efficient tool to reach these goals, as the reliable estimation of sales and the tendency can improve business and commercial strategies. At the organizational level, estimation of sales is an important input to many decision activities in some functional areas such as marketing, purchases, and or course sales, as well as finance and accounting. Thus, the ability of Spare Parts Sales Manager for heavy trucks to estimate the probable sales quantity in short term could help to improve the customer satisfaction, sales revenue and the efficiency of delivery. [12]

There are many developed estimating models in the world which are designed for different items, for example, it exists a multivariate repeat-sales model for estimating house price indices that is able to separately control for the effects of age and time, as well as other assets with changing attributes in the construction of price indices; the commonly used Case-Shiller repeat-sales model defines price change solely as a function of the difference in an asset´s price level between two points of time. The age adjusting price index not only reflects both the time and age effects, but is capable of incorporating different paths of depreciation. The empirical results indicate that controlling for the price effect of age in the age-adjusting index leads to a significantly different index than estimated by the traditional Case–Shiller approach.[7]

Other models of forecasting for new-released and nonlinear sales trend products, use high correlations between short and long term accumulated sales within similar products groups, and provide a prediction of long term forecast using the sales result of the product´s very early release. For practical use, this kind of model is designed to deal with the following three points: accuracy; timing of forecast release and the broad coverage of items.[8]

In the automotive industry there are some developed models, as Multi-step sales forecasting based on structural relationship identification, in which the empirical analysis indicates that automobile sales at segment levels have a long-run equilibrium relationship (co-integration) with identified

economic indicators; a vector error correction model (VECM) of multi-segment automobile sales was estimated based on impulse response functions to quantify long-term impact of these economic indicators: Consumer Price Index (CPI), Unemployment Rate, Gas Prices and Housing Starts. So, in this model is presented a structural relationship identification methodology that uses a battery of statistical unit root, weakly exogeneity, Granger-causality and co-integration tests in order to identify the dynamic couplings among automobile sales and economic indicators. [9]

There is a sales forecasting problem in the retail industry based on early sales, and nn effective multivariate intelligent decision-making (MID) model is developed to provide effective forecasts for this problem, by integrating a data preparation and preprocessing module, a harmony search-wrapper-based variable selection (HWVS) module and a multivariate intelligent forecaster (MIF) module. The HWVS module selects out the optimal input variable subset from given candidate inputs as the inputs of MIF. The MIF is established to model the relationship between the selected input variables and the sales volumes of retail products, and then utilized to forecast the sales volumes of retail products. Many experiments were conducted to validate the proposed MID model in terms of extensive typical sales datasets from real-world retail industry. Experimental results show that it is statistically significant that the proposed MID model can generate much better forecasts than extreme learning machine-based model and generalized linear model do. [10]

As it was mentioned, predictors, forecasting models are developed for different items and some studies combine variable selection method and support vector regression (SVR) to construct a hybrid sales forecasting model for computer products; in this case in order to evaluate the feasibility and performance of this approach, it compiles the weekly sales data of five computer products including Notebook (NB), Liquid Crystal Display (LCD), Main Board (MB), Hard Disk (HD) and Display Card (DC). [11]

Car industry cannot be isolated from the internet; such is the case of Google wherefrom using online search data, multivariate models and economic variables have been used for forecasting German car sales, in which Bayesian VAR model was performed and the results show rather well for all car brands and for short-and-medium-term forecasts. [13]

Some research have demonstrate that using sales proxies derived from a calibrated truncated log-normal distribution function can be used to produce realistic estimates of market average product prices, and product attributes during a forecast period where distribution function parameters are assumed to not change overtime, in which was used a straightforward method to produce an accurate approximation of sales volume using sales rank for refrigerators, freezers,, and clothes washers. [14]

As it can see, sales forecasting plays an important role in Business intelligence because can be designed methodologies and techniques used for acquiring and transforming raw data into structured and valuable information for analytic purposes. Estimating or Forecasting are the process of making predictions about the tendency of the future based on past and present data. Sales forecasting uses historical sales data, in association with products characteristics and peculiarities to predict short-term or long-term future performance, and it can be used to derive sound financial and business plans. [15]

Finally is important to consider the human side, i.e. the salespeople or sales representative, otherwise any sales estimating model has no reason to be created. Academics and managers would certainly agree that successfully leading salespeople is an activity of utmost importance for the success of any firm; through direction and coaching, sales leaders set goals for salespeople, align their activities with the objectives of the organization, and motivate them to perform well. To date, leadership in the sales domain has been approached as a top-down phenomenon. Consequently, the knowledge of whether salespeople can enact strategies to regulate their behaviors and self-lead is largely limited. But a research by arguing that self-leadership, and particularly thought self-leadership (TSL) strategy, is a viable construct that holds much promise for both academic research and practice. [16]

In the era of globalization, local knowledge is essential to firms ‘success, however, such knowledge is often difficult and expensive to acquire in foreign markets. From a buyer's perspective, sales rep`s who possess knowledge and insight into international markets are of great value, but sometimes cultural distance between buyer country and seller country alters the effectiveness of the trust-commitment building processes. When cultural distance between the two countries is low, the sales rep's capability trust is more important in building rep-owned commitment. In contrast, when cultural distance is high, the sales rep's benevolence trust is more important in developing rep-owned commitment. [17]

An intrinsically purpose of this estimating model is to visualizing the amount of purchase based on estimated sales; research indicates most sales strategies are effective but not all the time and there are two reasons of this lack of effectiveness: one of them is, strategies ignore the purchasing function, which is very important in business market; researchers estimate that approximately 80% of an organization's costs go through the purchasing department and the second, extant sales strategies do not recognize the different purchasing situations that can result from the different stages in the evolution of the purchasing function. [18]

III.BAYESIAN NETWORKS

Bayesian networks (BNs) are a formalism that in recent years has demonstrated its potential as a model of knowledge representation with uncertainty; this formalism originated as a contribution of different fields of research: decision-making theory, statistics and artificial intelligence. The successes of numerous applications in various fields such as medicine, information retrieval, artificial vision, information fusion, agriculture, etc., support this formalism. [4]

Bayesian networks represent the qualitative knowledge of the model through an acyclic directed graph, this knowledge is articulated in the definition of relations of independence, dependence between the variables that compose the model; these relationships range from total independence to a

functional dependence between model variables. In addition, they not only qualitatively model knowledge but also numerically express the "force" of relationships between variables; this quantitative part of the model is usually specified by probability distributions as a measure of the belief we have about the relationships between model variables.

Formally, a Bayesian network is a tuple B=(G,θ), where G is the graph and θ Is the set of probability distributions $P(X_i|Pa(X_i$ for each variable from i=1 to *n*, Pa(X_i) show the father of the X_i variable in the *G* graph.

To represent the quantitative part of the network model, the probability theory is used. In probabilistic environments, a probability distribution P can be considered a dependency model using the following relation:

$$I(X,Y|Z) \leftrightarrow P(X|YZ) = P(X|Z) \quad (1)$$

Where *X,Y,Z* are subsets of variables from the model, the sentence *I*(.,.|.) is interpreted as a conditional independence relationship, the above expression would read as "X is conditionally independent of Y, known Z", the codification of the conditional independence relations expressed in a BN make the joint probability distribution can be stored in a much more efficient and local way to each of the variables of the model, this means that:

$$P(X_i, \ldots\ldots X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i \quad (2)$$

IV. LEAST SQUARE METHOD

It is a procedure that allows estimating the parameters of any linear model and can be illustrated by simply fitting a line to a set of points. This procedure to fit a line passing through a set of n points that is, we want the differences between the observed values and the corresponding points in the adjusted line to be "small" in a general sense.

A convenient way of achieving this and providing estimators with good properties is to minimize the sum of the squares of the vertical deviations from the adjusted line.(Wackerly, 2010)

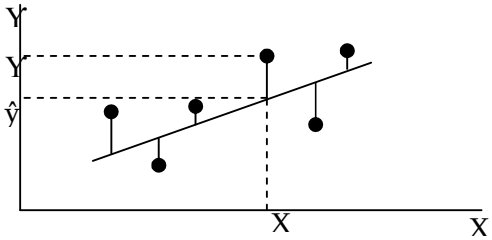

Fig.1,adjusting a line passing through a set of points.
Source: Wackerly, 2010

Therefore, a point interpolation can be performed through this method, by calculating the equation of the line of those involved variables, concluding in the following equations:

$$m = \frac{n\sum_{i=1}^{n} f(x_i)x_i - \left(\sum_{i=1}^{n} f(x_i)\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (3)$$

$$b = \frac{\left(\sum_{i=1}^{n} f(x_i)\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} f(x_i)x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (4)$$

Where n is the number of samples

The values in the equations are replaced to obtain m (slope) and b (ordered to the origin)

$$y = mx + b \quad (5)$$

The equation 5 represents the equation of the line.

V. DEVELOPMENT

As it was mentioned before, this model is for all those who have a managerial or strategic position where there is a need to have an estimate or trend of sale in the sale of truck parts in Mexico and the information for this purpose is based on five dealers of an important OEM in Mexico which is dedicated to sale and distribution of spare parts for trucks.

But the model is not only for those kind of dealers, is also for small dealers, such as independent dealers and to be clearer in this classification, in next figure it can see the structure of both kind of dealers that the model can be used for.

Fig. 2, OEM Dealer Organizational Structure
Source: Flores, 2015

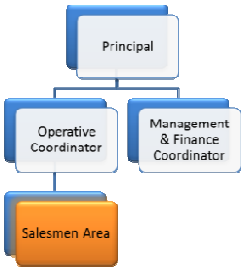

Fig. 3, Independent Dealer Organizational Structure
Source: Antonio, 2013

The spare parts distributors generate a significant industrial change based on a feedback in regards to necessary changes such as: consulting, development of new solutions, methodology of administration and quality; These are market-focused and go hand in hand with the global economy that has shifted product orientation to market orientation. For the operation of a distributor of marketing of spare parts is required of good outside salesman who permanently visit their customers, raise and / or generate the need for the customer to create new orders on a constant basis and preferably increasing amounts of the managers that watch permanently the accounts receivable of its routes avoiding the generation of bad accounts.

This is the reason for this research, it is a support of the sales area, generating and visualizing the trend of the sale in the short term to support the decision making in the stockpiling of sufficient inventory in store to support the projected sale and avoid lost sales and of course achieve the objectives established.

The proposed model consists of the following steps, first determine the variables that are considered to elaborate the sales predictor model, these variables can be seen in Table 1, for each of these variables an adjustment equation is determined, in this case it uses least squares and an equation of the line.

After calculating the adjustment equations, it can simulate a data that is not yet counted taking into account the statistical history, calculate the values of y for the 7 variables in use,

After that, these values are fed to a Bayesian network with the data previously fed and calculate the estimated sales for that date considering the estimate of the 7 chosen variables.

Table 1. Variables evaluated

# Var	Variable	Periodicity
0	Daily sale of truck spare parts	Daily
1	Monthly sale of motor units of national origin	Monthly
2	Vehicle fleet	By year
3	Monthly Production of Trucks and Integral Buses	Monthly
4	Monthly Inflation	Monthly
5	Gross domestic product (GDP)	Quarterly
6	Exchange rate	week

The period evaluated is between 2013 and 2016

VI. CALCULATIONOF THE EQUATION OF THE LINE FOR THE INVOLVED VARIABLES

1. Calculation of the equation of the line for the variable: Daily sale of truck spare parts.

Here we calculate the equation of the line by least squares using a line for approximation, for the variable: daily sale of truck spare parts. That in the future for an easier handling will be called Var_{sales},

n = 1224 (days)

$$m = \frac{(1224*16963880395559)-(2173771679*749700)}{(1224*612005100)-(749700*749700)} = \frac{(2.07638\,E^{+15})-(1.62968\,E^{+15})}{(7.48094\,E^{+11})-(5.6205\,E^{+11})}$$

$$m = \frac{(4.46701\,E^{+14})}{(1.87044\,E^{+11})} = 2388.218647$$

$$b = \frac{(2173771679*612005100)-(749700*16963880395559)}{(1224*612005100)-(749700*749700)} = \frac{(1.33036\,E^{+15})-(1.27178\,E^{+15})}{(7.48094\,E^{+11})-(5.6205\,E^{+11})}$$

$$b = \frac{(5.85772\,E^{+13})}{(1.87044\,E^{+11})} = 313173.3331$$

Thus:

$$y = 2388.2186\,x + 313173.3331 \qquad (6)$$

Fig. 4. Scatter plot ofVar_0

In fig. 4, 1224 points are presented, which represent each point one day, that go from the year of 2013 to 2016, therefore the axis X represents the time.The Y axis represents the sale in millions of pesos, in this scatter plot we can see an accumulation near to Axis X, but its increasing to 2016 year, this means that sales are increasing, and the fig. 2, show a line with upward sloping, this line represent the equation 4.

However, it can be seen that as it is an approximation line there are errors, for example if we check values f(x) where x=1, and x represent the January 2, 2013, f(x)= 1'035,092.86 , but if we use the approximation function, to determine f (x) we have the following:

$f(x_1)$=2388.2186x+313173.3331
$f(x_1)$ = 2388.2183(1)+313173.3331
$f(x_1)$ = 315561.55

Therefore there is an error.

$error_{x1}$ = |1035092.86-315561.55|
$error_{x1}$ = | 719531.31|
$error_{x1}$ = 719531.31

To calculate the error of this percentage we uses rule of three, if 1035092.86 represent the 100%,then the question is what does 719531.31represent?the calculation is seen in equation 5.

$error_{x1\%}$=(1035092.86*100)/719531.31 (7)
$error_{x1\%}$=69.5%

With this data it could be said that the error is high and that the adjustment function is not adequate, but it must be remembered that the method calculates uses 1224 samples, for this reason the calculation of differences for each case is made by using A relational database manager and applying

the use of SQL, after performing this calculation the following results were obtained:

The total of sales with 1224 samples was: $2173771679.44
The difference accumulated with 1224 samples was: $760955434.97898

The fit function error using least squares was: 35 %,

This indicates that the general error is less than that of a sample.

As it can see has evaluated a line with the least squares, now using minima squares will use a parabola in order to see if it adapts better and is decreased in the error in the selected variable.

Thisresearch is oriented to estimate the V_{sales} ,taking into account six variables, so it is important to prove if the dispersion of the data conforms to a parabola, so considering the same samples and data, it calculates the corresponding quadratic equation and model it to observe if the data is adjust to this type of curve.

Least square parabola has the equation below:

$$Y = a_0 + a_1 X + a_2 X^2 \quad (8)$$

Constants a_0, a_1 and a_2 can be calculated by solving the next equations, which are named as normal equations of least square parabola.

$$\sum Y = a_0 N + a_1 \sum X + a_2 \sum X^2$$

$$\sum XY = a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3$$

$$\sum X^2 Y = a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4$$

So that, based on same data and samples, values are substituted in the system of equations:

$$2173772 = 1224a_0 + 749700a_1 + 612005100a_2 \quad (E_1)$$
$$1696388340 = 749700a_0 + 612005100a_1 + 5.6205E^{+11}a_2 \quad (E_2)$$
$$1.5798E^{+12} = 612005100a_0 + 5.6205E^{+11}a_1 + 5.0058E^{+14}a_2 (E_3)$$

Solving by Gauss method, it has:
$E_1 \Leftrightarrow E_1$
$E_2 \Leftrightarrow E_2 + (-612.5)E_1$

$$133143513.5 = 749700a_0 \quad 459191250a_1 \quad 3.7405E^{+11}a_2$$
$$1696388340 - 749700a_0 + 612005100a_1 + 5.6205E^{+11}a_2$$

Solving the system, it has:
$$364952886.1 = 152813850a_1 + 1.8719E^{+11}a_2 \quad (E_{21})$$

$E_3 \Leftrightarrow E_3 + (-500004.1)E_1$
$$-1.0868E^{+12} = -612005100a_0 - 3.7486E^{+11}a_1 - 3.0600E^{+14}a_2$$
$$1.5798E^{+12} = 612005100a_0 + 5.6205E^{+11}a_1 + 5.0058E^{+14}a_2$$

Solving the system, it has:
$$4.9297E^{+11} = 1.8719E^{+11}a_1 + 2.4457E^{+14}a_2 \quad (E_{31})$$

So,it has partially the following system of equations:

$$2173772 = 1224a_0 + 749700a_1 + 612005100a_2 (E_1)$$
$$364952886.1 = 152813850a_1 + 1.8719E^{+11}a_2 (E_{21})$$
$$4.9297E^{+11} = 1.8719E^{+11}a_1 + 2.4457E^{+14}a_2 (E_{31})$$

$E_1 \Leftrightarrow E_1$
$E_{21} \Leftrightarrow E_{21}$
$E_{31} \Leftrightarrow E_{31} + (-1225)E_{21}$

$$4.4706E^{+11} = 1.0719E^{+11}a_1 \quad 2.2931E^{+14}a_2$$
$$4.9297E^{+11} = 1.8719E^{+11}a_1 + 2.4457E^{+14}a_2$$

Solving
$$45904762060 = 1.5262E^{+13}a_2$$

$$a_2 = \frac{45904762060}{1.5262\,E^{+13}} = 0.00300763$$

The value of a_2 is replaced in E_{21} equation

$$364952886.1 = 152813850a_1 + 1.8719E^{+11}(0.00300763)$$
$$364952886.1 = 152813850a_1 + 563019468$$
$$364952886.1 - 563019468 = 152813850a_1$$
$$-198066582.2 = 152813850a_1$$

$$a_1 = \frac{-198066582.2}{152813850} = -1.2961$$

Finally, values of a_1 and a_2 are substituted in E_1equation

$$2173772 = 1224a_0 + 749700(-1.2961) + 612005100(0.00300763)$$

$$2173772 = 1224a_0 - 971708.4 + 1840685.74$$
$$2173772 = 1224a_0 + 868977.24$$
$$2173772 - 868977.24 = 1224a_0$$
$$1304794.44 = 1224a_0$$

$$a_0 = \frac{1304794.44}{1224} = 1066.00853$$

Therefore the quadratic equation is:

$$Y = 1066.00853 - 1.296129\,X + 0.00300763\,X^2 \quad (9)$$

2. Calculation of the equation of the line for the variable: Monthly sale of motor units of national origin

n = 48 (months)

$$m = \frac{(48*3166371) - (123149*1176)}{(48*38024) - (1176*1176)} = \frac{(151985808) - (147175224)}{(1825152) - (1382976)}$$

$$m = \frac{(4810584)}{(442176)} = 10.8793$$

$$b = \frac{(123149*38024) - (1176*3166371)}{(48*38024) - (1176*1176)} = \frac{(4758665576) - (3723652296)}{(1825152) - (1382976)}$$

$$b = \frac{(1035013280)}{(442176)} = 2340.7269$$

Thus:
$$y = 10.8793\,x + 2340.7269 \quad (10)$$

3. **Calculating of the equation of the line for the variable: Motor Vehicle fleet**

n = 5 (years)

$$m = \frac{(5*6203948) - (2014992*15)}{(5*55) - (15*15)} = \frac{(31019740) - (30224880)}{(275) - (225)}$$

$$m = \frac{(794860)}{(50)} = 15897.20$$

$$b = \frac{(2014992*55) - (15*6203948)}{(5*55) - (15*15)} = \frac{(110824560) - (93059220)}{(275) - (225)}$$

$$b = \frac{(17765340)}{(50)} = 355306.80$$

Thus:

$$y = 15897.20\,x + 355306.80 \quad (11)$$

4. **Calculating of the equation of the line for the variable: Monthly Production of Trucks and Integral Buses**

n = 48 (months)

$$m = \frac{(48*15381015) - (5969772*1176)}{(48*38024) - (1176*1176)} = \frac{(738288720) - (702045072)}{(1825152) - (1382976)}$$

$$m = \frac{(36243648)}{(442176)} = 819.6631$$

$$b = \frac{(5969772*38024) - (1176*15381015)}{(48*38024) - (1176*1176)} = \frac{(2.26995\,E^{+11}) - (1.80881\,E^{+11})}{(1825152) - (1382976)}$$

$$b = \frac{(46113872952)}{(442176)} = 104288.5027$$

Thus:

$$y = 819.6631\,x + 104288.5027 \quad (12)$$

5. **Calculating of the equation of the line for the variable: Monthly Inflation**

n = 48 (months)

$$m = \frac{(48*227.2254) - (12.3570*1176)}{(48*38024) - (1176*1176)} = \frac{(15706.9207) - (15707.9229)}{(1825152) - (1382976)}$$

$$m = \frac{(-1.1022868)}{(442176)} = -2.49287\,E^{-06}$$

$$b = \frac{(12395.0*38024) - (1176*12.357.2254)}{(48*38024) - (1176*1176)} = \frac{(50/889.51) - (54481/.10)}{(1825152) - (1382976)}$$

$$b = \frac{(122072.40)}{(442176)} = 0.278333$$

Thus:

$$y = 2.4920\,E^{-06}\,x + 0.270333 \quad (13)$$

6. **Calculating of the equation of the line for the variable: Gross domestic product (GDP)**

n = 16 (Quarterly)

$$m = \frac{(16*817.1724) - (34.2231*136)}{(16*1496) - (136*136)} = \frac{(5074.75) - (4654.84)}{(23936) - (18496)}$$

$$m = \frac{(420.417)}{(5440)} = 0.07728$$

$$b = \frac{(34.2231*1496) - (136*817.1724)}{(16*1496) - (136*136)} = \frac{(51197.769) - (42135.457)}{(23936) - (18496)}$$

$$b = \frac{(8062.312)}{(5440)} = 1.4820$$

Thus:

$$y = 0.07728\,x + 1.4820 \quad (14)$$

7. **Calculating of the equation of the line for the variable: Dollar weekly exchange rate**

n = 210 (weeks)

$$m = \frac{(215*366406.98) - (3191.7557*22155)}{(210*3109085) - (22155*22155)} = \frac{(76945467.52) - (70713347.53)}{(652907850) - (490844025)}$$

$$m = \frac{(6232119.99)}{(162063825)} = 0.0384547$$

$$b = \frac{(3191.7557*3109085) - (22155*366406.988)}{(210*3109085) - (22155*22155)} = \frac{(9923439770.513) - (8117746823.57)}{(652907850) - (490844025)}$$

$$b = \frac{(1805692946.96)}{(162063825)} = 11.1419$$

Thus:

$$y = 0.03845\,x + 11.1419 \quad (15)$$

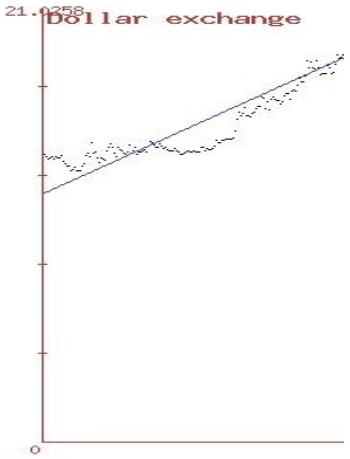

Fig. 5 Scatter plot of Var_6

In fig. 5, 210 points are presented, which represent each point one week, in the interval [2013,…,2016], therefore the axis X represents the time.The Y axis represents the dollar exchange rate, Mexican pesos by American Dollar, the fig. X, show a line with upward sloping, this line represent the equation 13.

n: 210
Standard deviation or σ = 2.4613059377258

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{210}\sum_{i=1}^{210} x_i = 15.198836$$

However it is observed that there are 46 samples above the standard deviation and 18 below

Max(x_i)= 21.0258
Min(x_i)= 12.1617

Summaries are now calculated.

$$\sum_{i=1}^{n=210} x_i = 22155$$

$$\sum_{i=1}^{n=210} f(x_i) = 3191.7557$$

$$\sum_{i=1}^{n=210} x_i^2 = 3109085$$

$$\sum_{i=1}^{n=210} x * f(x)_i - 366106.0882$$

m=0.038454726022294; b=11.141863071315

Thus
y=0.038454726022294 x + 11.141863071315 (16)

The difference accumulated with 210 samples was: 134.22795657174

The fit function error using least squares was: 4.2%

It can also be verified that the manual result of equation 15 is the same as that calculated per program presented in equation 16, are the same.

Equations 10 to 15, which serve as predictors, are taken to determine possible values, from a date that is not available, to prove this will be tested sales as of July 1, 2017

Since the x-axis represents the time and each day represents a unit, then if we want to determine the interpolated value of July 1, 2017 we take the days of the beginning of the year to that date and add to the sample of 1224 History of previous years

The day of this date is determined using MySQL with the script:

selectdayofyear('2017-07-01') from dual;

Then
x_1 = 48+6
x_1=54

y_1= 10.8793 x_1 + 2340.7269
$y_1 = 10.8793(54) + 2340.7269$
y_1 = 2928.2091

Follow the same process for the other variables
$y_2 = 15897.20(5) + 355306.80$
y_2 = 434792.8
$y_3 = 819.6631(54) + 104288.5027$
y_3 = 148550.3101
$y_4 = 2.4928\,E^{-06}(54) + 0.278333$
y_4= 0.27846
$y_5 = 0.07723\,(18) + 1.482$
y5 = 2.8730
$y_6 = 0.03845\,(234) + 11.1419$
y_6 =20.1392

These are the estimated values for this date, however, it will be necessary to consider other options of estimation, especially for the exchange rate of pesos per US dollars, since as of June 19, 2017, the exchange rate is 19.1 pesos per American dollar

VII. BAYESIAN CALCULUS

Bayesian network application for daily sales forecast given the attributes or variables:

1. Monthly sale of motor units of national origin
2. Vehicle fleet
3. Monthly Production of Trucks and Integral Buses
4. Monthly Inflation
5. Gross domestic product (GDP)
6. Dollar weekly exchange rate

Historical data of these variables are considered for the period from 2013 to 2016.The table 2 header contains the number of the variable being calculated

Table 2. Historical data sample

Date	1	2	3	4	5	6	Sales
2013-12-28	3168	381250	87195	0.57	1.13	13.09	517007.41
2013-12-30	3168	381250	87195	0.57	1.13	13.09	855608.69
2013-12-31	3168	381250	87195	0.57	1.13	13.09	1052221.29
2014-01-02	1725	395552	121520	0.89	2.27	14.76	639521.30
2014-01-03	1725	395552	121520	0.89	2.27	14.76	1244369.92
2014-01-04	1725	395552	121520	0.89	2.27	14.76	339877.37
2014-01-06	1725	395552	117574	0.89	2.27	13.14	1010040.32

Step 1

From the set of data of 1224 registers, the total probabilities are calculated.

C={639521.30}
X={sale of motor units of national origin = 1725; fleet= 395552; Production of Trucks and Integral Buses=121520; Inflation=0.89; Gross domestic product=2.27; Dollar weekly exchange rate=13.14}

Now we calculate P(639521.30|x).

Within the sales options there is a set of numerical values in the period 2013 to 2016 that meet the following condition C={9043.11 <= x <=16420364.38}

Table 3. Sample of sale of truck spare parts.

Date	Amount
2016-07-17	9043.11
2016-07-31	36225.73
2016-03-21	43133.37
…	…
2016-11-30	15362499.6
2016-06-30	15635459.9
2016-09-30	15824455.1
2016-10-31	16420364.4

As can be seen in Table 3, the values for C are quantities that are not groupable by the value because the 1224 registers have different values, then we have that *count*(C) = 1224, where the *count* function count the number of elements of set C.

Reason why it is not possible to continue with this data, to solve this it is proposed to create a table of segments to group them, for it has to take care of two aspects:

- How many segments are created
- Segment Size

For this first version it was decided to create 3 segments based on the arithmetic mean and the standard deviation, next the calculations are presented.

n: 1224

The average is calculated.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{1224}\sum_{i=1}^{1224} x_i = 1775957.254$$

The variance is calculated.

σ^2 =2086992674357.8

Thus standard deviation:

σ = 1444642.749

Then 3 segments are created as shown below:

Segment 1: $\{0 < x < \bar{x} - \sigma\}$
Segment 2: $\{\bar{x} - \sigma < x < \bar{x} + \sigma\}$
Segment 3: $\{\bar{x} - \sigma < x < \bar{x} + \sigma < x < \infty\}$

Substituting values

Segment 1: {0 < x < 331314.505}
Segment 2: {331314.505 < x < 3220600.003}
Segment 3: {3220600.003 < x < …}

Considering the historical values are as follows in terms of number of records in the segments:

Segment 1: 24 registers
Segment 2: 1075 registers
Segment 3: 125 registers

As you can see most historical records are concentrated in segment 2.

Thus
C={Segment 1, segment 2, segment 3}

Step 2.

Calculation of frequency tables from the history

The data is obtained using the SQL code:

selectvta_fecha, sum(vta_cantidad) from ventas_camiones WHERE year(vta_fecha) between 2013 AND 2016 group by vta_fecha order by sum(vta_cantidad);

Table 4. Frequency table for attribute: Monthly sale of motor units of national origin

		sale of truck spare parts		
		segment 1	segment 2	segment 3
sale of motor units	0-1950	3/24	73/1075	0/125
	1951-2000	0/24	0/1075	0/125
	2000 and more	21/24	1002/1075	125/125

In Table 4, you can see 3 columns and 3 rows, where the columns represent the segments of the daily sales amount variable, and the rows represent the domestic sales variable, the first element before the diagonal represents the summation Sales in that segment and the value after the diagonal represents the total of domestic automotive sold within the first segment of daily sales of parts of trucks, the other columns follow the same criteria of presentation.

The conditional SQL for row 1 of the table 4:

WHERE bay_vta_kw< 331314.505;
WHERE bay_vta_kw< 331314.505 AND bay_vta_unit< 1950
WHERE bay_vta_kw>= 331314.505 AND bay_vta_kw< 3220600.03 AND bay_vta_unit< 1950

WHERE bay_vta_kw>= 3220600.03 AND bay_vta_unit< 1950

Then the frequency tables for the other five variables are calculated, but here the calculation is not presented, but the same procedure is followed.

Step 3.

The posterior probability is calculated for the elements of the class C={Segment 1, segment 2, segment 3}, from the Bayesian network:

$$P(h|O) = \frac{P(O|h) * P(h)}{P(O)}$$

Where

h - Is the hypothesis.
O - Observations.
P(h|O) y P(O|h) - Conditional probabilities.

Since P (0 | h) is known as the likelihood of the hypothesis h has produced the set of observations O, it can also be expressed as follows

It can also be expressed as follows:

$$P(C|A) - \frac{P(A|C) * P(C)}{P(A)}$$

$$P(c_1, \dots c_k|a, \dots, a_n) = \frac{P(a_1, \dots, a_n|c_1, \dots, c_k) * P(c_1, \dots, c_k)}{P(a_1, \dots, a_n)}$$

Where C is the class, can take k possible values $\{c_1, \dots, c_k\}$ and $A=\{a_1, \dots, a_n\}$, set of n attributes.

Then in terms of classification it can be said that we are interested in finding the most plausible or more probable ci value a posteriori of the set C given the attributes. $a_1, \dots,$

A posteriori probability is also known as **M**aximum **AP**osteriori(MAP)[4], can be expressed as:

$$c_{MAP} = \arg\max P(c|a_1, \dots, a_n)$$

$$c_{MAP} = \underset{c \in C}{\arg max} \frac{P(a_1, \dots, a_n|c_1) * P(c_1)}{P(a_1, \dots, a_n)}$$

$$C_{MAP} = \arg\max P(c) \prod_{i=1} P(A_i|c)$$

C = {Segment 1}

Theprobability is calculated of C=Segment 1, given the attributes *x* provided by the user, from the frequency tables as shown in Table 4.

C=Segment 1

P(x| Segment 1) =
P(Var_1= 1951-2000|Segment 1)*
P(Var_2=Range2| Segment 1) *
P(Var_3=Range3| Segment 1) *
P(Var_4 =Range4| Segment 1) *
P(Var_5 =Range5| Segment 1) *
P(Var_6 = Range6| Segment 1)

$$C_{MAP} = \arg\max P(c) \prod_{i=1} P(A_i|c)$$

Then calculate:
$P(x|Segment\ 1)P(Segment\ 1)$

Then calculate: P(x| Segment 2), P(x|Segment2) P(Segment 2); P(x| Segment 3), P(x|Segment3) P(Segment 3)

Step 4.
Now calculate the total probability of the class C_1=Segment 1; C_2=Segment 2; C_3 = Segment 3.

$$P(x) = \sum_{i=0}^{n} P(C_i|x)P(C_i)$$

$$P(x) = P(Segment\ 1|x)P(Segment\ 1) + P(Segment\ 2|x)P(Segment\ 2) + P(Segment\ 2|x)P(Segment\ 2)$$

Step 5.
As we already have the necessary calculations are replaced in equation 1.

$$P(Segment\ 1|x) = \frac{P(x|Segment\ 1)}{P(x)}$$

$$P(Segment\ 2|x) = \frac{P(x|Segment\ 2)}{P(x)}$$

$$P(Segment\ 3|x) = \frac{P(x|Segment\ 3)}{P(x)}$$

After obtaining the three probabilities (Segment n | x), the highest value will be taken as feasible, and will therefore be the prediction.

VIII. CONCLUSIONS

Although the choice of interpolation of values obeys a straight line, for the six variables chosen to evaluate, it can also be concluded that for some cases the interpolated value is bit far from what should be, as is the case of the exchange rate of American dollars for pesos, which is necessary to continue working to improve the prediction model, but this could only be concluded after calculating the values and seeing the results.

On the side of the Bayesian networks, it was observed that it is necessary to find a better algorithm to create groups of values or clusters of values, with the purpose for grouping better and therefore not having so many segments, because more clusters or groups is the more complex the calculation.

It is also possible to conclude that the whole calculation process must be programmed with the fastest calculation, for this research work some calculations were done manually, but for the case of the predictors of the Bayesian networks, it was simply not feasible by the amount of clusters generated.

ACKNOWLEDGMENT

We like to express sincere appreciation and deep gratitude to all participants in this work. Authors kindly acknowledge financial and computational infrastructure support from Instituto Politécnico Nacional to facilitate the development of this work.

REFERENCES

1. Dennis Wackerly, Estadística Matemática con Aplicaciones, Cengage Learning, pp. 563-576, 2010.
2. Murray R. Spiegel, Estadística,McGraw Hill, 2009.

3. John E. Freund, Estadística Matemática con Aplicaciones, Pearson Educación, 2000.
4. José Hernández Orallo, Ma. José Ramírez Quintana, César Ferri Ramírez, Introducción a la Minería de Datos, Pearson, Prentice Hall, PP-259, México, 2008.
5. Juan Manuel Flores, Desarrollo de un Método Sistémico para la mejora en el proceso de ventas del área de repuestos, IPN Tesis, 2015
6. Gabriela Antonio, Modelo Viable de una empresa de comercialización de equipo industrial, IPN Tesis, 2013
7. Roger E. Cannaday, Henry J. Munneke, Tyler T. Yang, A multivariate repeat-sales model for estimating house price indices, Journal of Urban Economics, 2004
8. Kenji Tanaka, A sales forecasting model for new-released and nonlinear sales trend products, Expert Systems with Applications, An International Journal, 2010
9. Akkarapol Sa-ngasoongsong, Satish T.S. Bukkapatnam, Jaebeom Kim, Parameshwaran S. Iyer, R.P. Suresh, Multi-step sales forecasting in automotive industry based on structural relationship identification, Production Economics Journal, 2012
10. Z.X. Guo, W.K. Wong, Min Li, A multivariate intelligent decision-making model for retail sales forecasting, Decision Support Systems, Journal of Urban Economics, 2013
11. Chi-Jie Lu, Sales forecasting of computer products based on variable selection scheme and support vector regression, Neurocomputing Journal, 2013
12. Goodness C. Aye, Mehmet Balcilar, Rangan Gupta, Anandamayee Majumdar, Forecasting aggregate retail sales: The case of South Africa, Production Economics Journal, 2014
13. Dean Fantazzini, Zhamal Toktamysova, Forecasting German car sales using Google data and multivariate models, Production Economics Journal, 2015
14. Robert Touzani, Robert Van Buskirk, Estimating Sales and sales market share from sales rank data for customer appliances, Physica A, Statistical Mechanics and its applications Journal, 2016
15. F. Jiménez, G. Sánchez, J.M. García, G. Sciavicco, L. Miralles, Multi-objective evolutionary feature selection for online sales forecasting, Neurocomputing Journal, 2016
16. Nikolaos G. Panagopoulos, Jessica Ogilvie, Can salespeople lead themselves? Thought self-leadership strategies and their influence on sales performance, Industrial Marketing Management Journal, 2014
17. Flora F. Gu, Jeff Jianfeng Wang, Danny T. Wang, The role of sales representative in cross-cultural business-to-business relationships, Industrial Marketing Management Journal, 2016
18. Bert Paesbrugghe, Deva Rangarajan, Arun Sharma, NiladriSyam, SubhashJha, Purchasing-driven sales: Matching sales strategies to the evolution of the purchasing function, Industrial Marketing Management Journal, 2016