Improving Forecasting Accuracy to Reduce Variability of Customer Service Level

Mónica Balderas, Andrea Araiza, Mariana Peña & Bernardo Villarreal
Universidad de Monterrey
San Pedro Garza García, N.L., México 66238
monica.balderas@udem.edu, andrea.araiqa@udem.edu, mariana.penar@udem.edu, bernardo.villarreal@udem.edu

Abstract

This work describes the efforts of the leading Mexican convenience store company to look for opportunities to reduce a high level of customer service variability at the store echelon. The company was having high customer service variability originating that 42% of the time, on average, stores were out of stock. After an analysis of the structure of the distribution network and the actual forecasting and inventory management schemes, the authors found that the main cause behind the deficient performance was poor demand forecasting accuracy and incomplete inventory management scheme structures. Therefore, the improvement initiatives designed were developed to select correct or adequate forecasting procedures for the different demand patterns and the modification of the structure of the inventory schemes. Initial results are obtained from a pilot study carried out in all SKU’s of all the stores of the Puerto Vallarta plaza of the company. The resulting forecasting mean squared error (MSE) was decreased significantly in the order of 43%. The firm estimates a reduction of about 21% of the level of inventories and of 22 percentual points of customer service variability at the retailing stores due to the improvement in forecasting accuracy.

Keywords
Customer service variability, inventory management, retail echelon, demand forecasting

1. Introduction

An important sector in the Mexican economy where demand forecasting and inventory management are key tools for success is the retailing sector. The Mexican convenience store industry is rapidly growing along with the evolution of Mexican society. Mexican young families are changing customs, habits and roles. Now, both: husband and wife work and stay most of the day outside home. Time has become a very valuable asset to manage. This new environment has favored the emergence of the convenience stores (C-S). Location, fast response and 24/7 time availability have become key characteristics for the success of this format. Mexican C-S sector was positioned number 11 among the first 15 biggest world markets in year 2014 with total sales of 8500 million dollars. This level of sales was generated by 17,450 stores established throughout Mexico. The leading company in this sector contributed with 12,853 stores and a market share of 88% in that year. This firm will be called “The Main” hereafter.

One of the greatest challenges of convenience stores to be competitive refers to demand forecasting and inventory management. These are fundamental to maintain adequate levels of product availability and insure customer satisfaction. This aspect represents an important weakness for “The Main”. The company was experiencing high levels of variability of its customer service response.

This work has the purpose of describing the efforts of “The Main” to decrease the firm’s customer service variability at the store echelon level. The document is structured as follows. The first section presents an introduction and general context. Second section describes a summary of bibliographic research relevant to the problem of interest.
The following section provides a description of the general methodology followed to treat the problem. Then, the application of this methodology is given in the fourth section, followed by the fifth section of results and conclusions.

2. Review of concepts and methodology for forecasting

In an important study of retail out-of-stocks (OOS) provided in Corsten et al. (2003), it is stated that “availability of products is the new battleground in the fast-moving consumer goods industry”. The concept of stockouts in the retail sector is not new. Progressive Grocer (1968a,b) provide the first major study on how grocery customers reacted to stockouts. Research about this situation has been very intense. These studies identify five main reactions by consumers to a stockout in store:

1. They buy the item at another store.
2. They delay ordering or purchasing the item (postpone purchase at the same store).
3. They do not purchase the item (a lost sale).
4. They substitute the same brand (different size or type).
5. They substitute for another brand (brand switching).

An excellent study of grocery customers in London developed by Schary et al. (1979) revealed that 48 percent of the customers select to shop in another place when faced with a stockout. According to studies developed by the Mexican leading convenience store company, 40% of the customers would switch to purchase a substitute item and the rest would shop in another place when face with the same condition. Research described by IGD (2003) shows that 65 percent of UK consumers looking for a specific grocery item will adopt one of the first three reactions, thus not buying in that particular store on that occasion if a stock-out occurs. In 1979, the figure from the study described in Aastrup et al. (2010) was 78 percent. Compared with the more general results provided by Corsten et al. (2003), the figure for the UK is high compared with other markets where the average is 31 percent.

2.1 The causes of retail out-of-stocks

An excellent retail reference model was developed by Aastrup et al. (2010). This is very useful to categorize main causes of stockouts. The works described in Gruen et al. (2008), McKinnon et al., (2007), Fernie et al. (2008) and Aastrup et al. (2009) are used to summarize the most significant causes. These can be classified as follows:

- Pre-store causes: These are related to direct suppliers or the retailer’s distribution center. Some of the most important are: Deficient general planning and communications; Deficient warehousing procedures; forecasting problems; inaccurate inventory transaction recording; unreliable transportation and others.

- Instore causes: These occur after the inbound replenishment process to the store has finished. Among the most common are: store ordering problems; deficient manual inventory adjustments; items damage; problems in the process of moving items from the back-store to the correct space on the retail shelf and; promotion-caused stockouts in the stores.

Most OOS situations occur at the store level according to Corsten et al. (2003), primarily through ordering and replenishment practices. About 35 percent of OOS problems occur with shelf replenishment in the store, 30 percent occur during the inbound delivery to the distribution center of the company, 15 percent from the regional distribution center (RDC) to the store and 15 percent is due to inventory accuracy problems.

2.2 Approaches to improve on-shelf availability

Two of the main approaches suggested for improving OSA are provided by Corsten et al. (2003) and by the ECR UK (2004). Corsten et al. (2003) developed an approach to address the causes of OOS. This is based upon the achievement of an improvement of process responsiveness, operational accuracy and incentive alignment. Process response improvements were related to assortment planning and space allocation; ordering systems, inventory control and store flow replenishment. Operational accuracy initiatives are focused upon the accuracy of inventory
levels and the ability to measure and identify the On-Stock Availability level (OSA). Finally, incentive alignment, is about scheduling staff to improve shelf filling in addition to optimizing overall management objectives.

In their report, the ECR UK (2004) recommended a combination of processes and approaches to increase the level of OSA. The study identified seven “levers” that can be used to improve OSA. These seven levers are: measurement “levers” which need “managerial attention”; replenishment and in store execution, namely merchandising; inventory accuracy; promotional management and ordering systems. These levers have subsequently formed the basis for several OSA improvement strategies.

**2.3 Forecasting concepts and schemes**

Forecasting is a prediction of future events used for planning purposes that has to do with the estimation of the value of a variable (or set of variables) at some future point in time. Krajewsk, et al. (1998) see forecast as a prediction of future events used for planning purpose and this is needed to aid in determining what resources are needed, scheduling existing resources and acquiring additional resources. Forecasting is a task that can be accomplished with the support of qualitative and quantitative forecasting methods (Jacobs et al. 2011). For our case, quantitative models are our concern. In particular those forecasting models constructed based on time series.

The appropriate forecasting procedure or scheme to use depends very much on the behavior of the demand pattern. Two types of demand patterns are of importance in this paper. The first demand pattern refers to the main elements that form it; randomness, average, trend, seasonality and cycles (Heizer, J., et al. 2013; Chase, R. et al. 2004). Thus, the forecasting procedure to be used will depend on the existence of the previous elements in the pattern of demand. The second type of demand pattern regards the degree of variability on the size and the time of occurrence of demand (Williams 1984; Syntetos et al. 2005). The accuracy of a forecasting method for a particular product depends on characteristics exhibited by the product’s demand pattern. Consequently, demand time series are sometimes divided into several discrete categories in order to assign the best forecasting method. The idea of categorizing demand patterns initially appeared in Williams (1984), who studied the classification of products by demand type, stock control policies for different categories of products, and methods of forecasting demand for the different categories of products.

A new approach to this problem was suggested by Syntetos et al. (2005) (to be called SBC hereafter). SBC categorize demand based on the expected mean square error of each forecasting method under some assumptions. They compare the method suggested by Croston (1972) (hereafter CRO) and a bias-adjusted version of Croston’s method due to Syntetos et al. (1999) and hereafter referred to as SBA. From this comparison they propose the four discrete categories of demand shown in Figure 1 which they label ‘erratic’, ‘lumpy’, ‘smooth’ and ‘intermittent’. The four quadrants are uniquely specified by two parameters \( p \) and \( v \), where \( p \) is the average inter-demand interval and \( v \) is the squared coefficient of variation of the demand when it occurs. The threshold values defining the quadrants are given as \( p = 1.32 \) and \( v = 0.49 \) respectively. Both CRO and SBA use a smoothing constant for producing exponentially smoothed estimates of positive demands. They also both use the parameter \( p \) to denote the average inter-demand interval.

![Figure 1. Illustration of intermittent demand pattern categories](image-url)
An interesting mechanism for improving the forecasting performance was recently provided by Nikolopoulos et al. (2011). This is called “An Aggregate-Disaggregate Intermittent Demand Approach (ADIDA)”. This approach is mainly based upon a non-overlapping temporal aggregation of demand in higher level time buckets (say, from days into weeks for example). The result of the application of this tool is the reduction of zero-demand periods, having a new demand pattern showing a smoother behavior with lower variability levels. Nikolopoulos et al. (2011) empirically showed that ADIDA can result in lowered forecasting errors and characterized it as a forecasting method self-improving mechanism.

Spithourakis et al. (2011) illustrates that ADIDA also performs well in many cases for non-intermittent demand data. Similarly, cumulative demand of intermittent series will exhibit significantly less intermittence, any aggregate series would be expected to have a considerable reduced coefficient of variation compared to the original series. In other words, aggregation may smooth out time series randomness. The most important benefit of ADIDA is that it is an inexpensive scheme for managers to estimate highly accurate forecasts. Therefore, ADIDA can be regarded as a cost efficient and universally implementable forecast accuracy-improving mechanism. For the special case of using the resulting forecasts for inventory management, Nikolopoulos et al. (2011) proposed a managerial heuristic for the aggregation level, setting it equal to an item's lead plus its review time. This is particularly useful for periodic review inventory management systems.

2.4 Inventory management concepts and schemes

The academic literature that treats production-inventory management has its origins since early 1900’s with the works by Harris (1915), Emery (1954) and Gorden, et al. (1956) among others. These original works have been further enriched by numerous researchers since then. For excellent descriptions of relevant results, the reader is referred to Silver et al. (1998) and Hadley et al. (2012). Most of the initial literature has considered customer demand as stationary. Parameters required for continuous or periodic review management systems are estimated based on a sample of demand history under this assumption. Further, they are not updated in the future to consider possible demand pattern changes making them obsolete and inadequate for an efficient inventory management. The previous resulting inventory control policies were called “static” by Babai et al. (2009). This type of policies is not recommended if the actual product demand pattern contains additional elements such as trend, seasonality or cycles. The approach to determine static control policies is called by Babai et al. (2009) the “inventory consumption-based approach”. On the other hand, there are product demand patterns that exhibit volatile conditions with trend and/or seasonality included. Under this environment, advanced demand information is very valuable and required to insure an efficient inventory management. This information can be provided by a forecasting system or through customer orders. Inventory management parameters can be updated using this information in a dynamic manner. The previous approach used to define inventory control policies is called the “future-requirements based approach” in Babai et al. (2009). The two main policies recommended in the literature (Silver et al. 1998, Hadley et al. 2012) are the re-order point (r, Q) and the order-up-to-level (T, S).

Retailers are often dealing with an inventory replenishment environment in which deliveries are periodically (based on a delivery schedule per store), replenishment quantities are an integer multiple of a fixed case pack size, sales follow a weekly pattern with peak sales on Friday and Saturday and shelf space per SKU is limited. In practice, a complete inventory policy must include case pack size (order quantity) and shelf capacity as decision variables. In fact, this policy should be developed considering an alignment among the previous variables. However, such alignment and optimal decisions are difficult to achieve due to the existing fragmented approach to inventory decision making in retailing. Although retail operations may be responsible for setting reorder points, shelf space decisions are often made by the retailer’s merchandising and/or marketing organizations. Furthermore, case pack size is generally determined unilaterally by the supplier on the basis of pallet dimensions, truck trailer dimensions, and packing machine capabilities. As each party attempts to optimize the decision variable under their control, their efforts will not be completely effective in achieving full alignment (optimality) among case pack size (order quantity), shelf space, and reorder point.

Donselaar et al. (2008) present a comparison of two inventory replenishment strategies in a retail environment; The Full Service (FS) strategy and the Efficient Full Service strategy. Both strategies represent an effort to integrate the concepts of reorder point, case pack size and shelf space into a single inventory policy. In the Efficient Full Service strategy, if at a review period the inventory position, IP, is strictly below the reorder level s, we order the maximum

© IEOM Society International
number of case packs, Q, such that the inventory position (IP) after ordering is less than or equal to the shelf capacity V. Unless this IP is still below s, i.e., the shelf is not large enough to accommodate all units, then we order as many case packs as needed to bring the inventory position after reordering to (or just above) s. In summary: if at a review period IP is strictly less than s, the order quantity, q, becomes:

$$q = \max \left\{ \left( \frac{V - IP}{Q} \right) \cdot Q, \left( \frac{s - IP}{Q} \right) \cdot Q, 0 \right\} \quad \text{if } IP < s \quad (1)$$

The reorder level, s, is equal to the average forecasted demand during the review period, R, and delivery period, L, plus the safety stock, ss, for a given predetermined service level.

3. Description of Methodology

After reviewing both approaches previously described in section 2.2, the team responsible for the initiative decided to follow the ECR UK model as a broad conceptual guideline for identifying the most important areas for improving potential lost sales.

The study conducted in “The Main” followed the following steps; The initial work is focused on the analysis of the company’s context including its supply chain structure, product catalogue, forecast characteristics and the like. At the same time, the team included a task of focusing where the problem was occurring. This last action would facilitate the analysis of finding the main causes of the customer service variability. The following task corresponds to the identification of the demand patterns and types of items included in the study. Then, current forecasting and inventory management procedures used, together with their actual performance are analyzed. The fourth step was designed to include an initial simulation study of the impact of using ADIDA for a sample of fast and slow-moving items sold by “The Main”. Afterwards, and considering the achievement of positive results in the simulation, a pilot program would be devised for a significant sample of products sold at the Puerto Vallarta plaza was carried out. The results of this program were the basis for convincing the management of the firm to pursue the implementation of the mechanism for the rest of the store network nationally during year 2018.

4. Application of Methodology

The leading convenience store company, “The Main”, has grown in an impressive manner during the last decade. The supply chain structure includes 16 Regional Distribution Centers (RDC) that service a total of 15,225 stores as of the end of year 2016 with a market share of 88% in Mexico. “The Main” has a two-echelon divergent inventory system, also known as the one warehouse and N-retailers inventory system (Axäter, 2000). The company uses simple exponential smoothing for forecasting daily demand for all A and B items. For all C – F items, “The Main” utilizes simple moving averages to forecast daily demand. All inventories managed in “The Main” use a periodic review system with a review period of one week.

4.1 Focusing the analysis on problem stores and products

The starting phase of the detailed analysis consisted of defining the “plazas” (groups of stores), with the highest level of customer service variability levels. As shown in Figure 2, the plazas of Puerto Vallarta, Tampico and Mexicali were the most important in this regard. The team decided to select Puerto Vallarta for further work.

Next, the study continued to identify the product families with the greatest percentage of lost sales during year 2018. As presented in Figure 3, the families of general merchandise, edible groceries and hygiene and personal health items are the most important.
Given the history of daily sales for the items included in the selected families from January to September of 2018, these were used to make an ABC analysis and plotted in a Syntetos matrix to identify their demand characteristics. The first study resulted that 78% of the items are classified as C items and 20% of them as B items. Under the Syntetos scheme, it was found that 87% of the items had intermittent demand and 13% were lumpy (see Figure 4).
4.2 Description of current forecasting and inventory management schemes

The actual inventory management scheme used by “The Main” for all items is described as follows. The inventory management system is a periodic review order up to system. The review period is one week, and the replenishment delivery frequency is daily. This system includes an order up to maximum quantity, M, calculated as the average demand during the review period plus the delivery response time and safety stock. Inventory management parameters can be updated using this information in a dynamic manner. The previous approach used to define inventory control policies is called the “future-requirements based approach” in Babai et al. (2009). Further, even though the company uses the periodic review scheme for all the items, only A items consider safety stock for the estimation of the M parameter.

From Figure 4, it was concluded that C – F items have a demand pattern that is mostly lumpy and erratic and B items are mostly erratic. For these type of items Syntetos et al. (2005) recommend the application of the SBA forecasting procedure for this type of demand pattern. The company is currently using simple exponential smoothing and moving average procedures for this task. Hence, it seems that changing the method would improve forecasting precision. Furthermore, the estimation of at least the Safety Stock will be better with a good chance of being lower. In addition, as previously mentioned in section 2.3, the mechanism recommended by Nikolopoulos et al. (2011) called “An Aggregate-Disaggregate Intermittent Demand Approach (ADIDA)” can also be used to improve forecasting performance. The tool contemplates four steps; (1) Gather original data; (2) Apply a non-overlapping temporal aggregation at an aggregation level, A; (3) Extrapolate the aggregate time series by means of a forecasting method, F; and; (4) Disaggregate aggregate forecasts back to the original time scale via a disaggregation algorithm, D.

Given the previous description of the forecasting and inventory management procedures, the identified potential improvement initiatives consist of replacing the forecasting procedure considering bigger periods of time (moving from days to weeks) and the determination of safety stock for managing inventories.

4.3 Impact of the implementation of initiatives on customer service level and inventory

Thus, in our case, the first step carried out was the change of the forecasting procedures and the application of the ADIDA process. A significant quantity of items (B, and C – F) were selected from the Puerto Vallarta plaza for running simulation studies and a pilot program of one month. The aggregation period chosen was equal to the review period of one week. Figure 5 presents the impact of aggregating daily demand to weekly demand. As shown, the new demand patterns for items B and C-F present a significant trend towards the smooth category. Under this new category, it will be possible to apply the CRO technique or other procedures such as simple exponential smoothing and moving averages.
The simulation studies and pilot program were designed to consider the possibility of using demand aggregation and various forecasting procedures such as simple exponential smoothing, moving averages and Croston.

The best result obtained from the previous initiatives consists of the application of simple exponential smoothing and ADIDA, resulting in the highest Mean Absolute Deviation (MAD) reduction. For the pilot program, the SKUs with the highest percentage of lost sales were selected (400 SKUs). For this, starting on week $n$, the sales for the past four weeks were considered to forecast $n+1$ using simple exponential smoothing and ADIDA, resulting in variable $x$. The $x$ value is then shared with the replenishment team for them to consider it as the reorder point for each given SKU. Week by week, results were measured by comparing sales vs. forecast. This pilot program was implemented for 3 weeks. Figure 6 illustrates the impact of these efforts on the percentage of lost sale for the worst five stores of Puerto Vallarta plaza. The percentage of lost sale was reduced 58% on average.

The second phase of the implementation effort, the simulation study, consisted on modifying the inventory management scheme for B and C-F items to include safety stock. Together with this modification, it was also necessary to test the normality of the probability distribution function of item demands. After a statistical analysis of goodness of fit, it was determined that most of the items had a log normal demand distribution. From this, the same SKUs as the pilot program were selected (400 SKUs). The basis for the simulation were the sales for seven registered weeks. The forecast was calculated for each week, using simple exponential smoothing and ADIDA plus the calculated safety stock. This resulted in six variable $x$'s that represented the stock required per SKU to be supplied each week. The Mean Absolute Deviation (MAD) indicator was recalculated for each of the seven weeks. The average percentage of lost sale per store in Puerto Vallarta plaza decreased an additional 21% (see Figure 7).
5. Conclusions

The level of competitiveness of the retailing sector is dependent on the level of product availability at the store. Under these circumstances, the effectiveness of the forecasting and inventory management schemes becomes important.

The case of study treated in this paper deals with the improvement of the level of precision of demand forecasting at the store level. The company was dealing with the problem of having excessive level of customer service variability in its stores. An exhaustive analysis of the contribution to this excess was developed and found that forecasting performance for inventory management purposes could be improved. Thus, an initiative based on the application of the ADIDA (Nikolopoulos et al. 2011) methodology and a change of forecasting procedure could be implemented for these purposes. Additionally, it was also identified the need to modify the periodic inventory management scheme to include safety stock for B and C-F items. As described in section 4, the impact of implementing the previously mentioned efforts was very positive. Simulation studies and pilot program actions estimate a total reduction of customer service variability of the order of 79%.

Based on the previous positive results, the company decided to make a full implementation of the initiatives in Puerto Vallarta. The following step agreed by the operations management consisted of deploying a similar analysis and design scheme for the rest of the plazas in Mexico. The application described in this work provides additional evidence of the benefits of implementing ADIDA to improve forecasting accuracy for intermittent demand items.

References

Emery, J.C., Some Theoretical Aspect of Inventory Control, Master’s Thesis in Industrial Management, M.I.T, 1954.


**Biographies**

**Mónica Balderas** is a MAGNA CUM LAUDE Industrial Engineer graduated from Universidad de Monterrey (UDEM). She has been part of different projects, such as applying Lean Thinking principles to processes for a global Oil and Gas industry to improve productivity and efficiency. She also participated in the improvement of operations related to safety procedures for the local water and sewage company. Monica is currently working as project manager for British American Tobacco, working on the development of new products for different end markets.

**Mariana Peña** is a CUM LAUDE Industrial Engineer graduated from Universidad de Monterrey (UDEM). She has been part of different projects such as the improvement of the safety in the operations of a water and sewage service company. She also applied Lean Thinking principles for improving the productivity of different processes in a company that supplies equipment for the global oil and gas industry. Mariana also worked in the Human Resources area of a industry 4.0 startup. Mariana is currently working as a consulting analyst in a management consulting firm.

**Andrea Araiza** is an Industrial Engineer just graduated from Universidad de Monterrey (UDEM). She has participated on several projects such as the Improvement of the OEE in an assembly line at a metal manufacturing company. Also implemented several projects related to the levels of scrap in a company that manufactures catalytic converters and escape systems. Currently, she has started to work at a multinational biotechnology product development company. Andrea is a member of the IIE Society.

**Bernardo Villarreal** is a full professor of the Department of Engineering of the Universidad de Monterrey. He holds a PhD and an MSc of Industrial Engineering from SUNY at Buffalo. He has 20 years of professional experience in strategic planning in several Mexican companies. He has taught for 20 years courses on industrial engineering and logistics in the Universidad de Monterrey, ITESM and Universidad Autónoma de Nuevo León. He has made several publications in journals such as *Mathematical Programming, JOTA, JMMA, European Journal of Industrial Engineering, International Journal of Industrial Engineering, Production Planning and Control, Industrial Management and Data Systems* and the *Transportation Journal*. He is currently a member of the IIE, INFORMS, POMS, and the Council of Logistics Management.