Noname manuscript No.
(will be inserted by the editor)

A Data Mining Approach to Forecast Behavior

Nihat Altintas · Michael Trick

November 2010

Abstract This study presents a data mining analysis of customer forecasting patterns of multiple customers (auto manufacturers) from a large auto parts supplier. We consider a manufacturing environment in which forecasts of future orders are used as inputs for a series of decisions. We define the complexities that are captured from our data set, developing the daily flow analysis to obtain accuracy ratios of forecasts as a performance measure for customers. We also demonstrate the application of some recent developments in clustering and pattern recognition analysis which can have a significant impact on the performance analysis of customers.

Keywords Data Mining · Automotive Industry · Forecasting

1 Introduction

The rapid expansion of computer resources creates the potential to bring business intelligence into decision-support systems of supply chains. With the increased availability of enterprise-wide databases, the amount of data collected is growing at a tremendous rate. The company databases are full of lots of hidden information that creates an opportunity for the companies to better understand and analyze their operations by looking at the data. This becomes more critical for supply chains with many players and several constraints. One objective of supply chain management is the reduction or elimination of all

F. Author
Credit Suisse, 11 Madison Avenue, New York, NY 10010
Tel.: +1-347-2202250
E-mail: nihat.altintas@credit-suisse.com

S. Author
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213
Tel.: +1-412-268-3697
E-mail: trick@cmu.edu

activities that do not add value and concentrating on factors that maximize value and productivity. That requires a joint effort for Collaborative Planning, Forecasting, and Procurement(CPFR) among the players in the supply chain. It is important to improve the relationship among all participants in the supply chain through jointly managed planning and shared information. The quality and the intensity of the information exchange demands a strong commitment to cooperation from the participating organizations. The planning and forecasting both require intensive information exchange in all levels of relationships. Unsatisfactory relationships between the parties lead to inefficient information flow in the supply chain.

In our analysis, we look at a real life example from automotive industry of this information sharing and provide analytical methods how to use information more efficiently to improve the operations. Automotive industry demonstrates itself as a good example for potential improvements with better data analysis due to the strong relationship structure among the players, complex structure of products and geographic factors. It is not possible to make high level generalizations without mining the data and exploring it in a systematic way.

We consider a customer order prediction model in which forecasts of orders at some future date are used as input to a series of inventory planning decisions. We are analyzing the orders that are received by a supplier who produces parts for auto manufacturers. Multiple customers (auto manufacturers), requesting multiple parts, place preliminary orders which are estimates for actual orders starting from six months before the due date. Customers can update their orders as the due date approaches. The supplier guarantees very high service levels as stipulated in the contract with the customers. Since capacity is limited and there is lead time for production, the supplier may fail to fulfill orders. Therefore, production planning and capacity decisions have to be made carefully in order to achieve high service levels. As the production quantities are being committed so far in advance, it is very difficult to predict final quantities of each part desired. Our goal is to improve the supplier's operations through a better understanding of her customers' ordering behavior.

We provide a methodology that can be used in order to analyze the forecast behavior of manufacturers. In the first step of the analysis, manufacturers' forecast data are transformed into a format that can be used as input for further analysis. Orders are replaced by daily flows in order to predict the daily requirements of the manufacturers and overcome any complexities due to ordering problems such as order-splitting, order-combining and changes in due date. In the second step of the analysis, we use data mining techniques such as clustering and projection methods in order to visualize forecast behavior of the manufacturers.

In our analysis, we introduce different order complexities that have not been discussed in literature. Analytical models that assume similar ordering behaviors can be used to obtain policies that improve the supplier's operations.Another important point is the lack of empirical research in the area

of supply chain management for order behavior. Our results provide strong empirical support for models that assume different forecast behaviors in literature. In our analysis, we show that customers are consistent with their forecast behavior over time. We provide examples of order over-estimation and under-estimation. There are models in literature that explain inflation of order sizes under different assumptions (Cachon and Lariviere 1999a, 1999b). However, the research to provide real life examples for ordering behavior is not well documented. Our research helps to understand underlying factors behind forecast performance.

2 Environment Analysis

2.1 Automotive Suppliers

Automotive industry has transitioned to more buyer driven model over the last 10 years. As auto manufacturers focus on the customer side, production and engineering move more towards suppliers. Suppliers become more important part of valuation creation in automotive supply chain. Problems faced at major auto manufacturers due to parts problems resulted more cost and lost brand values for the companies. Therefore, as documented in a survey by A.T. Kearney, it is critical for auto supplier to build close relationships with their customers and have strong networking abilities to be successful in the industry.

Automotive manufacturers have highly geographically dispersed operations. In 2009, 61 million vehicles were manufactured in 39 different countries(OIAC 2009). Therefore, it is critical for the automotive suppliers to serve an international network of manufactures. Although globalization is a well-established trend among the industry's biggest players, many of the largest suppliers still remain heavily rooted in their home markets. High customization of products and geographically dispersed operations require an efficient supply chain management for auto parts. Decreasing margins in industry and high technology requirements result in very demanding auto manufacturers. With the consolidation of the industry due to the recent economic slow down requires suppliers to be more cost efficient and more responsive to the customers at the same time.

Decreasing margins in industry and high technology requirements result in very demanding auto manufacturers. For example, supplying Honda requires patience. In some cases, the auto manufacturer would talk with a potential supplier as long as two years before deciding to offer a contract. Larry Jutte, head of purchasing for Honda of America Manufacturing, Inc. comments about their search for suppliers in the North American market: "We began by trying to identify suppliers in North America that matched our core values. Some thought Honda's value system was more challenging, they thought it was unreasonable. There is no question that we were demanding about quality, cost and delivery, we wanted to know a heckuva lot more about a supplier than

most other auto companies at that time." Therefore, a strong tie and collaboration is a key to lengthy and profitable relationships in the automotive industry.

In our analysis, we provide a problem from a major global auto parts supplier. The supplier has a process where he collects forecasts from the manufacturers starting from 6 months before the order delivery. The forecasts keep updated during the time before the delivery. It is critical to use this shared information efficiently to better to manage the supply chain with less cost and high service levels. Our objective is to understand the forecast behavior of customers by looking at historical data and provide meaningful and consistent patterns that help to predict to future demand. Automotive industry provides an interesting opportunity with several manufacturers and multiple parts to make inferences regarding the forecast behavior of companies.

2.2 Order Forecasts

Order forecasts are an essential part of collaboration when the capacity is limited and there is lead time for production. Forecast is the basis for the integration of a manufacturer to the production process. In CPFR, even sharing point of sales data is not enough for effective and efficient production planning. Sharing demand information alone will not guarantee uncertainty reduction in the system. One barrier is the time horizon. Since there is lead time and orders are placed in fixed epochs due to fixed costs, manufacturers should provide early forecasts.

There are two components of variation in the system. First, one is the noise which comes from the uncertainty of the nature. The current demand becomes a better predictor for the forecasted demand with the proximity to actual delivery. A forecast for tomorrow will be more accurate than one for next month. The latter is the manufacturer's bias in terms of over-estimating or under-estimating the order size. Manufacturers can improve their forecasting performance and can reduce the noise. However, as long as it is not incentive-compatible to submit true forecasts, manufacturers will transform their forecast, and truthful transmission of forecast is not possible. In our analysis, we provide empirical support for different forecast behaviors for the manufacturers.

2.3 Database Information

In our analysis, we are analyzing the forecasts and orders that are placed by multiple customers (auto manufacturers) to a supplier. The supplier is one of the biggest engine systems suppliers in the automotive industry. We provide an analysis for his "hot"-selling and high investment ($5.3 billion) new engine system. Multiple customers, requesting multiple engine parts, place preliminary orders that are estimates for actual orders starting from six months before the due date. Customers can update their orders as the due date approaches.

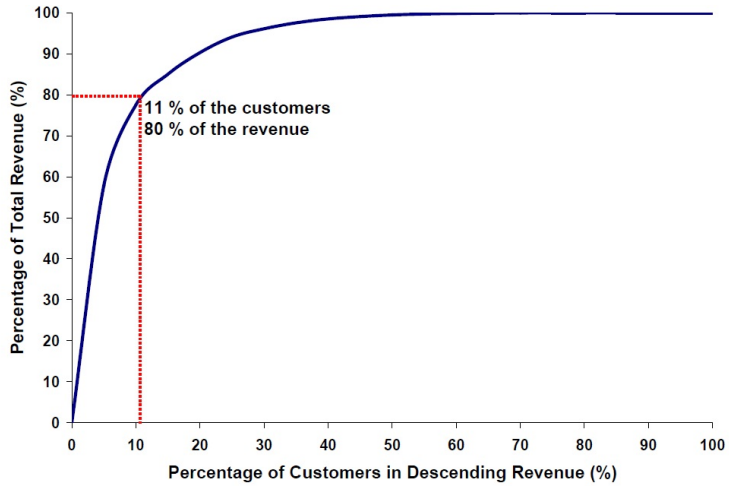

Fig. 1 11% of the customers accounts for 80% of the revenue

Our database consists of the orders placed by 497 customers for 35,551 different parts in years 2000 and 2001. The total number of orders is more than a million. Each part belongs to one of 214 product families. Parts in product families have similar functions and prices. An average price for a product family is also available in the data.

The attributes for different databases are as follows:

1. **order database:** customer, part, order date, due date, quantity, warehouse, factory information
2. **part database:** product family
3. **product family database:** average price for family

Pareto Result: Customers in our analysis are not identical in total dollar value of their orders. A widespread empirical result about income dispersion, known as Pareto result, implies that 20% of the customers will account for about 80% of the total demand. Ordering the customers in descending revenue in Figure (1), 11% (55 customers) of the customers represents 80% of the revenue.

Some of the customers represent $3-5\%$ of total revenue. Each of the top 31 customers' share in total revenue is higher than 1%. Therefore, by analyzing those 31 customers (just 0.6% of customers), we can explain 65% of the total revenue. For the next steps of our analysis, we separately analyze the major customers and provide statistically more significant results.

2.4 Objective

As part of CPFR, the supplier asks customers (auto manufacturers) to submit forecasts in order to start production in advance. However, a customer's

forecast accuracy can be very low and forecasts can be very misleading. Our goal is to define a framework for analyzing the forecast performance of the customers and to provide empirical support for different forecast behaviors. In our analysis, we explore and analyze large quantities of data to discover meaningful patterns and rules for the ordering process of customers. We try to explore the following questions:

a. Are classical forecasting techniques applicable? If not, is it possible to clean the forecasts in order to prepare the data for a deeper estimation analysis?
b. Is it possible to provide a quantitative representation of a customer's forecast performance?
c. Can we visualize the behavior of customers through time?
d. Can customers be assigned to groups? Do these groups represent significant commonality among different forecast behaviors?
e. How can we recognize different forecasting patterns? Is it possible to automate the process of anomaly detection from the data?
f. Are customers consistent with their forecast behavior? What is the general tendency in terms of over-forecasting or under-forecasting?

3 Literature Review

The management science literature has paid a great deal of attention to forecasting, so we only provide a description of major classes of models related to our analysis. Most of the literature is based on providing mathematical models for evolution of forecasts. Graves et al. (1986a, 1986b, 1998) and Heath and Jackson (1994) develop the Martingale Model of Forecast Evolution (MMFE) to model how demand forecasts of customers evolve in time. In MMFE, a forecaster generates a demand forecast for a single item in different periods. The forecast update errors, the difference between any two periods' forecast for a future date, are assumed to follow the Martingale Property: forecast update errors are independent, identically distributed, multivariate normal random variables with mean 0. Gullu (1996) , Graves et al. (1998), and Toktay and Wein (2001) use MMFE to model the evolution of the forecasts. Another approach is using Bayesian updates to incorporate new information as it becomes available . Scarf (1959) , Azoury (1985) and Lariviere and Porteus (1999) provide insights into the order evolution problem from Bayesian perspective. Chen et al. (2000) , Aviv (2001) and Aviv (2002) have suggested several other stylized theoretical models to study inventory planning when there are forecasts available for future demand. From the analysis of the customer forecast data, we capture complexities with order updates such as order-splitting, order-combining and shifts in due dates. The demand forecasts does not follow a regular pattern and the data without any transformation does not fit in the classical forecasting models.

Collaborative forecasting is critical for an efficient supply chain. Taylor and Xiao (2010) show that the manufacturer is hurt by the poor forecasting

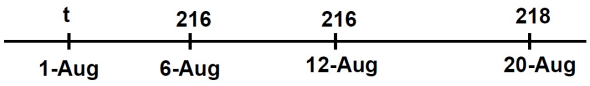

Fig. 2 A customer forecast at order date 1-Aug.

performance of a retailer. Aviv (2007) demonstrates that the benefit of collaborative forecasting is higher when the manufacturer has the largest explanatory power. Our analysis provided a framework to analyze the orders of a customer to better plan production for a manufacturer.

The empirical research about forecast behavior of customers is quite limited in literature. Terwiesch et al. (2003) considers the problem from a buyer's perspective. A buyer is placing orders from a set of equipment suppliers. The forecast performance of the buyer is measured according to the forecast volatility (continually updating the orders) and forecast inflations. They demonstrate that inflating forecasts and providing volatile orders damage the buyer's reputation and thereby lead to a lower level of service in the future. In our analysis, we looked at the problem from the supplier's perspective and provide analysis to understand the forecast behavior of the customers by looking at data.

4 Understanding Customer Forecasts

At each order date t, customer i update all his previous forecasts and can place some new forecasts for part j. Since we repeat the same analysis for all customers and parts, we drop the indices i and j for the rest of the analysis. Therefore, at order date t, the customer places forecasts for different due dates $S_t = (s_{1,t}, s_{2,t}, ...)$ where $s_{r,t}$ is the r^{th} due date of the order date t. Time between order dates and time between due dates should not necessarily follow a regular pattern. Each customer can provide a forecast update at any date. At t, the customer provides a set of forecasts $F_t = (f_{s_{1,t}}, f_{s_{2,t}}, f_{s_{2,t}}, ...)$ for all days in S_t. The forecast vector F_t is the most recent forecast information at time t and $f_{s_{1,t}}$ denotes $(s_{1,t} - t)$ days advance forecast. The most recent order date provides the valid forecast information for estimation. If there are no more updates for a due date, the forecast is treated as the final order. In the example in Figure (2), for t=1-Aug, the due date vector is S_t=(6-Aug, 12-Aug, 20-Aug) and the forecast vector for these particular due dates is $F_t = (216, 216, 218)$.

4.1 Order Definition

In our analysis, customers can modify both their due dates and forecast quantities at order dates. Changes in due dates make it difficult to track the updates of an order at different order dates. From the data set, we provide the major complexities that make classical forecasting tools inapplicable.

Example 1 12-Aug and 20-Aug orders in 1-Aug order splits into 12-Aug, 16-Aug, 20-Aug and 24-Aug orders in 8-Aug order.

1-Aug		**8-Aug**	
Due date	**Forecast**	**Due date**	**Quantity**
6-Aug	216		
12-Aug	216	12-Aug	107
		16-Aug	109
20-Aug	218	20-Aug	112
		24-Aug	106

4.1.1 Order-splitting

Some customers substitute their forecasts for a given due date with smaller batches as the due date gets closer. Customers tend to place large preliminary aggregate forecasts and then split them into smaller orders. We observe that a customer can place monthly forecasts six months in advance and change it into biweekly forecasts three months in advance and finally end up with weekly forecasts in the last month before the shipment. In Example 1, forecasts for due date 12-Aug and 20-Aug are replaced by 12-Aug, 16-Aug, 20-Aug and 24-Aug forecasts in 8-Aug order. When a forecast gets split, we lose track of this particular order. We have to make extra assumptions in order to obtain the actual order quantity for the order that gets split.

4.1.2 Order-Combining

Some customers combine their orders as the due date gets closer. The initial order no longer exists after combining. Therefore, there is no actual order quantity for that forecast. The accuracy of the order before combining is hard to derive without making any extra assumption about a customer's combining policy. In Table 2, forecasts for 12-Aug, 16-Aug, 20-Aug and 24-Aug orders from 1-Aug order are replaced by forecasts for 12-Aug and 20-Aug orders at 8-Aug.

Example 2 12-Aug, 16-Aug, 20-Aug and 24-Aug orders in 1-Aug order are combined into 12-Aug and 20-Aug orders in 8-Aug order.

1-Aug		**8-Aug**	
Due date	**Forecast**	**Due date**	**Quantity**
6-Aug	110		
8-Aug	124		
12-Aug	124	12-Aug	256
16-Aug	132		
20-Aug	112	20-Aug	252
24-Aug	140		

Example 3 8-Aug, 12-Aug, 16-Aug, 20-Aug and 24-Aug orders in 1-Aug are shifted by two days to 10-Aug, 14-Aug, 18-Aug, 22-Aug and 26-Aug in 8-Aug order..

1-Aug		**8-Aug**	
Due date	**Forecast**	**Due date**	**Quantity**
6-Aug	110		
8-Aug	124	10-Aug	120
12-Aug	124	14-Aug	128
16-Aug	132	18-Aug	134
20-Aug	112	22-Aug	120
24-Aug	140	26-Aug	134

4.1.3 Shifts in Due Dates

Customers can modify due dates of orders while updating their forecasts. We observe that it is difficult to keep track of an order when there is a change in its due date. Initial forecast starts six months before the shipment. Therefore, shifts in due dates are inevitable in a manufacturing environment with lots of uncertainty in production and demand sides. In Example (3), forecasts for 8-Aug, 12-Aug, 16-Aug, 20-Aug and 24-Aug orders are replaced by forecasts for 10-Aug, 14-Aug, 18-Aug, 22-Aug, and 26-Aug orders. In Example (3), there is a two-day shift in due dates and the customer is also adjusting his forecast quantity.

A New Order Evolution Model

In the MMFE framework, the order vector is updated at each order date. In our data set, the customer does not have to provide an order update at each order date t. The time between order dates with updates does not necessarily follow a pattern. In MMFE, the customer is assumed to provide forecasts at each order date for each period in the planning horizon. Customers can only update the quantities and cannot change the due date of an order in MMFE. However, in our analysis, the customers first choose the due dates S_t at order date t, and then provide the forecast vector F_t for these days in S_t. The customer does not provide forecasts for all days in the planning horizon. He submits forecasts only for the due dates in S_t. The number of days between the due dates in S_t does not necessarily stay constant . In our analysis, we also observe that customers can shift due dates. Moreover, order-splitting creates new due dates and order-combining reduces the number of due dates. Order-combining, order-splitting and shifts in due dates can happen simultaneously, which makes it difficult to track orders.

When an order gets split, combined or shifted, having order numbers does not solve the problem of finding an accuracy level for this order. In order to handle these complexities, we look at the periods in which the customer consumes these quantities. We assume that customers consider their daily production requirements while placing orders. A systematic approach is dis-aggregating the order quantities and work at daily flow level.

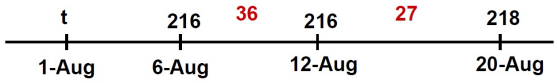

Fig. 3 Daily flow analysis for Example (4).

Example 4 At order date t=1-Aug, 216 is ordered for 6-Aug and it will get consumed between 6-Aug and 12-Aug (6 days), so the daily flow between 6-Aug and 11-Aug is $216/6 = 36$. Between 12-Aug and 19-Aug, the daily flow is $216/8 = 27$ following the same argument.

Order date	Due date	Forecast	Number of days	Daily flows
1-Aug	6-Aug	216	6	36
	12-Aug	216	8	27
	20-Aug	218		

4.2 Daily Flow Analysis

At each order day customers aggregate daily requirements and place orders for some due dates which are discrete points in the planning horizon. In our daily flow analysis, we try to predict the daily requirements of a customer based on the forecasts. There are different ways of estimating daily flows from orders. The greedy solution is assigning any order to the following days before the next due date. So, at order date, we divide the orders $f_{s_{k,t},t} \in F_t$ by $s_{k+1,t} - s_{k,t}$ and assign it evenly to the days between $s_{k+1,t}$ and $s_{k,t}$. We assume that $f_{s_{k,t},t}$ is consumed during those days. Other smoothing techniques can be considered with extra assumptions. In our analysis, we assign daily flows after a due date, unless the order is the latest due date of an order forecast vector F_t. Otherwise, we cannot find the days when the quantities ordered at the last days of an order date get consumed. Example (4) and Figure (3) demonstrate the assignment of daily flows for order date 1-Aug.

Handling order-splitting, order-combining and shifts in due dates is important for performance analysis of the customer. However, our operational objective is to understand how customers change their order quantities including the orders which do not have any complexities. We provide a procedure to find a quantitative representation for each customer's forecast performance.

1. **Finding the Accuracy Ratio of Forecasts**
   By running the daily flow analysis on the forecast update at order date t, we can obtain the daily forecasts ($\hat{f}_t^m$) for each day m in the planning horizon at order date t. By repeating the same daily flow analysis on the firm orders, we can generate the actual daily flows d_m for any day m. Having the actual daily flows and the daily forecasts, we can determine

the accuracy ratio of an r-day advance order. Different time windows can be considered for accuracy calculations. In our analysis we consider 30-day advance orders since most customers provide forecasts at least 30-days before the due date. The same analysis can be repeated for other r values. One major complexity in finding the accuracy level of a 30-day advance order is defining a 30-day advance order. It is not often the case that there exists a particular order which is placed exactly 30 days from the order date. One solution is comparing forecast daily flows and actual daily flows for the 30-day in advance. However, this is very sensitive to small shifts in due dates. A more robust solution is considering the total orders in a time window. From the data, we observe that customers provide orders on a weekly basis as due date gets closer. Therefore, we combine daily flows between 30 and 36 days (a week of orders) as a single order and compare it with the sum of actual daily flows for these 7 days. So the accuracy ratio ϕ_r^t of r-day advance order at order date t is

$$\phi_r^t = \frac{\sum_{n=r}^{r+6} \hat{f}_t^n}{\sum_{n=r}^{r+6} d_n}. \tag{1}$$

2. **Taking the Log of the Accuracy Ratios**
   It has been discussed in literature that for ratio-scaled data, using log transforms has more accurate results. (Armstrong (2000)) Hausman (1969) transformed the order evolution problem into a finite horizon sequential decision problem using a quasi-Markovian or Markovian model. He provides empirical support (not entirely) for ratios of successive forecasts being independent and having a Lognormal density function conditional on the change in the forecast. Therefore, we take the log of the accuracy ratios in our analysis in order to give the same weight to under- and over-estimation of the orders.
3. **Assigning Orders to Bins**
   In order to represent the behavior of a customer, we create a histogram for each customer by looking at 30-day advance orders. We assign orders to bins (different intervals in the histogram) by looking at the log value of the forecast's accuracy ratio. We consider 15 different bins with equal widths. There are three groups of bins, one true-estimator (accuracy ratio ranged from 80% to 125%), seven under-estimators (having ranges scaled being multiples of $\log 0.8$) and seven over-estimators (having ranges scaled being multiples of $\log 0.8$).
   For each customer, we start to fill in the bins as the daily flows are calculated. Based on the log value of the accuracy ratio, the appropriate customer bin is increased accordingly. The customers can replenish different parts with different quantities. Therefore it is reasonable to have an amount of increase which is equal to the dollar value of an order. After completing assignment of orders to bins for a customer, we normalize the

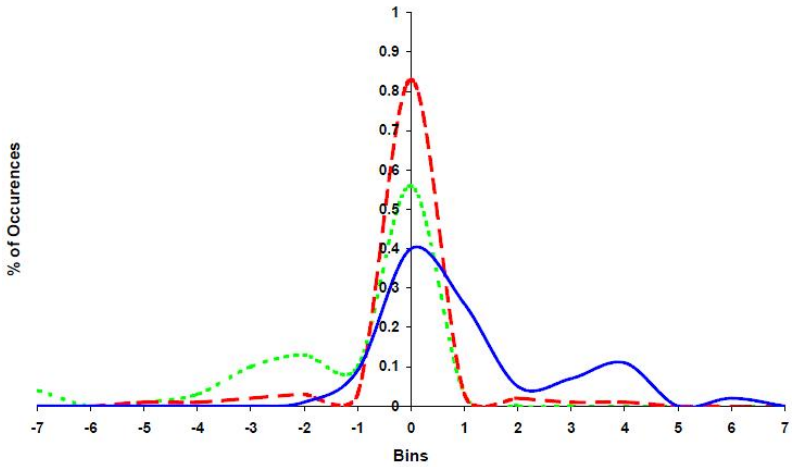

Fig. 4 The dashed curve corresponds to a customer who falls into the true-estimator bin 85% of the time. The solid curve exhibits the behavior of a customer who over-estimates the size of his orders most of the time. The dotted curve represents the ordering pattern of a customer who under-estimates the size of his actual orders.

histogram and find the relative frequency of each bin. So the value of each bin corresponds to the ratio of total dollar value of 30-day advance orders that fall into this accuracy level. We can use this 15-dimensional vector to mimic the distribution of a particular customer and input it as a performance measure in data mining. In Figure (4), resulting order distributions for 3 different customers are shown.

As a result of our analysis, we have a *customer order distribution* p_i^q for customer i and quarter q. We consider quarters of 2000 and 2001 in our analysis. Each p_i^q is a vector of size 15. The first 7 components of p_i^q(indexed from -7 to -1) corresponds to under-estimation values. In Figure (4), 0 on x axis which is the 8^{th} component of the vector p_i^q represents the probability of true-estimation and the last 7 components of p_i^q (indexed from 1 to 7) shows the over-estimation probabilities.

5 Characterizing Customer's Forecast

An important issue for a supplier with multiple customers is to provide performance benchmarks. Customer forecast data contain lots of variables which make it difficult to make comparisons among the customers. In Section 4, we provide the methodology to derive a quantitative representation of a customer's forecast performance. In this section we provide a distance metric in order to compare customer order distributions of customers. We then provide supervised and unsupervised clustering techniques that form performance groupings among customers.

Clustering heuristics assign each observation or object to a group. We can cluster the customers by looking at their order distributions. If a customer clusters with over-estimators or under-estimators, then he must be treated with care, on the grounds that there is a problem with his order process. If he clusters with true-estimators, that concern disappears. We can observe the evolution of a customer's forecast behavior by looking at his cluster for various quarters. We can call the cluster which includes the customer with 100% forecast accuracy as the ideal cluster, and explore the other clusters based on their distance from the ideal cluster.

In data mining, the main goal is to produce simplified descriptions and summaries of large data sets. As long as there are only two or three dimensions it is easy to visualize two- or three-dimensional graphs. However, as the dimensionality of the data gets larger, it gets difficult to plot a vector of relationships between different factors. Projecting high-dimensional data sets as points on a low display (usually 2-dimensional) is a one way of visualizing the data. In our analysis, we project the clusters on a two dimensional space.

A simple clustering is dividing customers into three groups: over-estimators, true-estimators and under-estimators. However, there are advanced clustering techniques which provide better insight. We apply two clustering techniques (K-Medoid Analysis and Self-Organizing Maps) in our analysis. For projection to two-dimensional display, we consider Sammon's Mapping that finds a mapping such that the distances between the image points of the data items remain similar to distances in the original metric. Sammon(1969) describes a nonlinear mapping algorithm which has been found to be highly effective in the analysis of multivariate data. The special feature in Sammon's Mapping is that errors are normalized by distances in original space. In our analysis, we consider data for different quarters of 2000 and 2001 for each customer. By looking at the trajectories of the customers on the mapping, customers can be informed about their performance in order to improve their ordering process and offer incentives.

5.1 Comparing Customer Performances

Customers can be represented in terms of proximity between their performance vector. In order to compare customer order distributions, we define a distance measure that provides a good approximation for the proximity between the customers.

The distance measure should be compatible with the problem characteristics. The vector of customer distribution(p_i^q) has some distinct properties. Each component corresponds to a discrete probability, and the sum of the mass probabilities adds up to one. The order of the vector is important and should be considered in the selection of the distance measure. The following distance measure is used in the rest of our analysis.

Definition 1 Having the customer probability vector p_i, a distance $d(i, j)$ between customers i and j can be defined as follows

$$d(i,j) = \sum_{k=-7}^{k=7} \| \sum_{t=-7}^{t=k} p_i^q(t) - \sum_{t=-7}^{t=k} p_j^q(t) \| \quad (2)$$

where $p_i^q(t)$ denotes the t^{th} component of vector p_i^q and $\| \ . \ \|$ stands for euclidean metric.

Our distance measure provides the distance between the cumulative probabilities. Since each distribution is an ordered vector, the distance measure should depend on the order of the vector. Our distance measure is robust to small errors in the calculation of the mass probabilities and puts more weight on the cumulative probability of any point. Also, the distance is not skewed to over-estimation or under-estimation. The distance between two order vectors stays the same if we transpose the order vectors. We use our distance measure to form customer clusters which have similar forecast behavior.

5.2 Customer Clustering Analysis

The goal of clustering analysis is to partition the observations into groups so that pair-wise dissimilarities between those assigned the same cluster tend to be smaller than those assigned in different clusters. Each observation is assigned to one and only one cluster. The objective of our clustering analysis is to describe forecast behaviors. Each cluster represents customers who have similar forecast behavior. By looking at the movement of customers between the clusters for different quarters, we can track the ordering pattern of a particular customer over time. If a customer shows up in the same cluster for all quarters, this means that the customer is consistent with his ordering behavior. Our clustering analysis also automates the process of setting performance benchmarks. Clustering also provides a dynamic measure of the ideal behavior for customers.

5.2.1 K-Medoid Analysis

We introduce K-Medoid clustering technique, which is a modified version of the well-known K-Means clustering technique. The only difference is having actual customers as the cluster centers. In our updated version of the heuristic, we try to find cluster centers which minimizes the maximum distance from any points in the cluster. Cluster centers are actual customers. The clustering technique is based on the assignment of each customer to the closest cluster center. The main issue in this clustering analysis is to map the customers according to a cluster center. If we take the sum of distances, then the customers can be well-spread between the clusters. However, we try to look at the case when a customer has an irregular behavior and is assigned to a cluster whose center has similar types of behavior. In order to get a better spread among the clusters, we suggest Kohonen Networks in the next section.

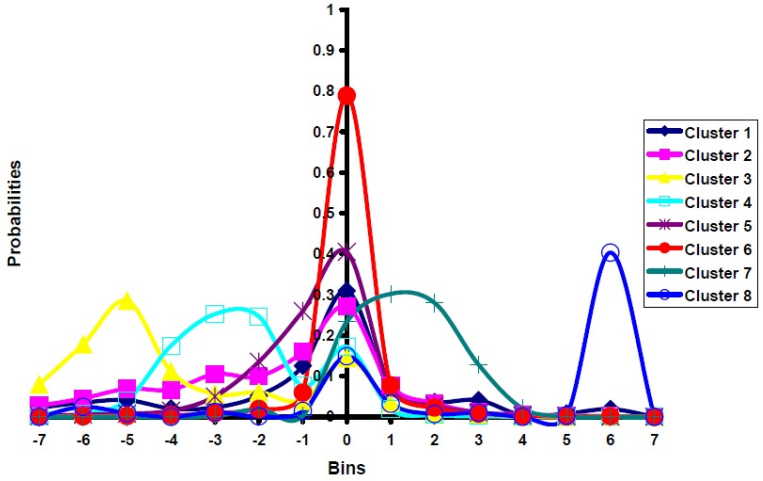

Fig. 5 Average customer distributions for each cluster for K-Medoid Analysis.

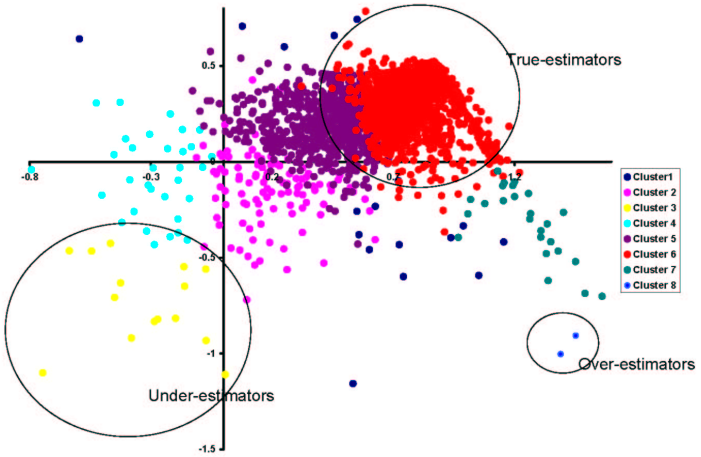

Fig. 6 K-Medoid Analysis for all customers. Each point represents a customer.

In K-Medoid Analysis, we try to automate the process of recognizing irregular customer behavior. If a customer tends to have a forecast behavior which is different from the other customers, he may become a cluster center and pulls other customers with similar behavior to his cluster. We do the clustering analysis for seven quarters of data of 2000 and 2001. Customers are assigned to eight different clusters. Each cluster represents a different customer behavior. In order to understand the properties of customers for each cluster, we plot the average customer distribution function for each cluster in Figure (5). Cluster 6 represents the customers whose forecasts are true estimates of the actual orders about 80 percent of the time. Ideal customers also

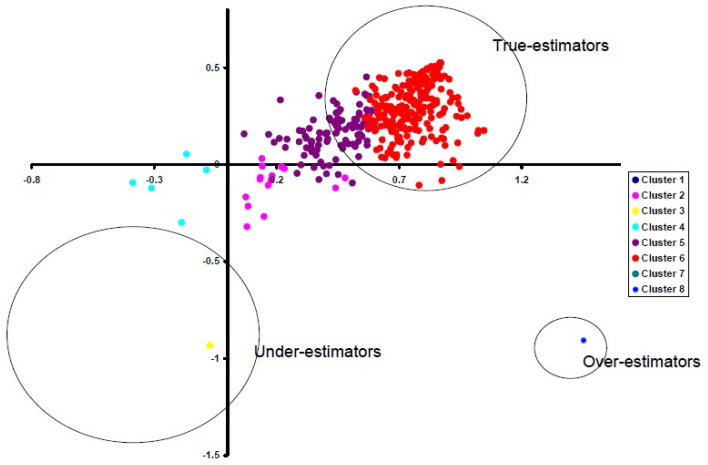

Fig. 7 K-Medoid Analysis for top 11% of the customers. Each point represents a customer.

fall into cluster 6. Clusters 7 and 8 can be considered as over-estimators and Cluster 3 and 4 can be considered as under-estimators with different levels. Clusters 1, 2 and 5 show more irregular patterns that are mixes of over-, true or under-estimation.

In Figure (6), the Sammon's Mapping of the K-Medoid clustering is shown. Cluster 6 (red points), which we call the **ideal cluster**, includes the customers who are close to the ideal customer. As we get far from the ideal cluster, we can observe different clusters. Different regions on the mapping define different behaviors. The lower-right part of the graph represents customers which are over-estimators (Clusters 7 and 8). Left of the y-axis represents the customers who are under-estimators. (Clusters 3 and 4). The other clusters (which have different behavioral patterns with different magnitudes) occupy different regions on the mapping.

In Figure (7), we observe the mapping for clusters for the top 11% of the customers who represents 80% of the total revenue. Most of the major customers fall into cluster 6 which contains ideal customer. We can still observe the other clusters with different behaviors. Therefore, our cluster analysis provides strong clusters.

5.2.2 Self-Organizing Maps (Kohonen Networks)

The Self-Organizing Maps (SOM) is an effective tool for the visualization of high-dimensional data. It converts the relationship of object in high-dimensional space into simple geometric relationship of their image points in a lower dimensional grid. SOM compresses information to display and produces some kind of abstraction. Kohonen (2001) described SOM as a nonlinear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements

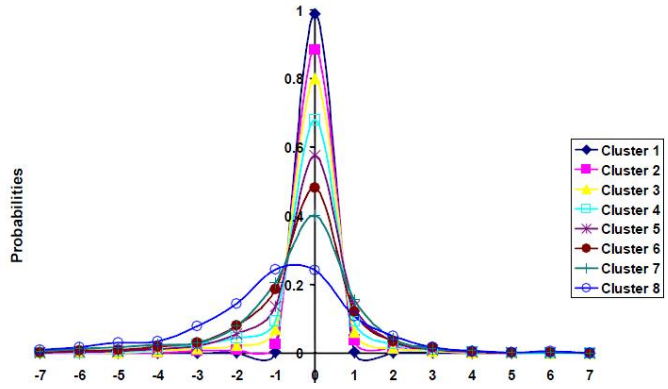

Fig. 8 Average customer distributions for each cluster for Kohonen Network Analysis.

of a regular, low-dimensional array. SOMs are also named as Kohonen Networks.

We consider a SOM with one-dimensional eight clusters. Since SOMs are topological maps, distance between the Kohonen clusters represents the level of dissimilarity. The cluster numbers in K-Medoid Analysis do not provide any information about similarities of clusters. Compared to K-Medoid Analysis, Kohonen Network provides a better spread among the clusters. Figure (8) provides the average customer distributions for different clusters. Cluster 1 contains the customers who truly estimate their orders more than 90% of the time. When we compare ideal cluster of Kohonen Analysis (Cluster 1) with the ideal cluster of K-Medoid Analysis (Cluster 6), we can observe that Kohonen Analysis has a small number of customers in ideal cluster with higher average true-estimation ratio than K-Medoid Analysis. Compared to K-Medoid Analysis, customer-quarter results between the clusters have more spread and so it is harder to recognize irregular forecast behaviors compared to K-Medoid Analysis. To capture irregular patterns, K-Medoid Analysis seems to provide better result. This is due to the fact that K-Medoid Analysis tries to minimize the maximum distance from cluster center, so any irregular customer-quarter result has to be close to a cluster center or become a cluster center himself in the output.

The Sammon's Mapping of Kohonen Analysis is a mountain-like shape. As we get far from the ideal customer, we observe points which have similar distances from the ideal customer, forming circular clusters. It is hard to recognize who are over-estimators or under-estimators. The advantage of Kohonen Analysis is having a better spread and not giving too much weight to single observations. K-Medoid Analysis is better for recognizing outlier behaviors. When we analyze the major customers, we still observe all of the clusters. (Figure 10)

Customers are Consistent:

In our analysis the clustering analysis provides significant clusters. In K-Medoid Analysis 36% stays in the same cluster for all quarters and 90% stays

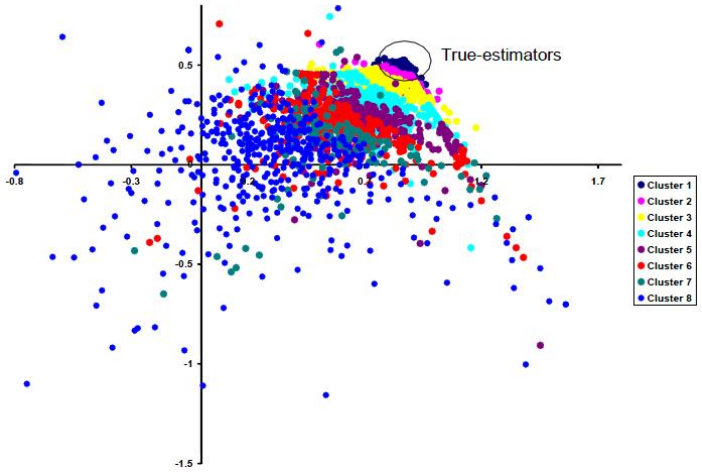

Fig. 9 Kohonen Network Analysis for all customers. Each point represents a customer.

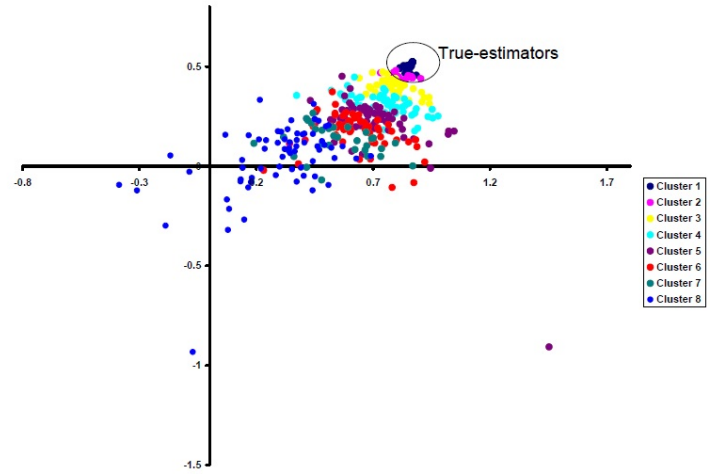

Fig. 10 Kohonen Network Analysis for top 11% of the customers who represent 80% of the total revenue. Each point represents a customer quarter result.

in the same cluster at least half of the quarters. In Kohonen Analysis 15% stays in the same cluster for all quarters and 56% stays in the same cluster at least half of the quarters. Since some of the clusters in K-Medoid Analysis have large number of customers, these clusters occupy a larger volume in the space. Therefore, the probability of staying in the same cluster is higher compared to Kohonen Analysis, where we have small clusters with equal size.

Reputation is Important:

In clustering analysis, we provide customer groups which provides irregular behaviors. When we take the averages of all customer distributions, we obtain a smoother behavior in Figure 11. On overall, true-estimation (indexed 0) seems

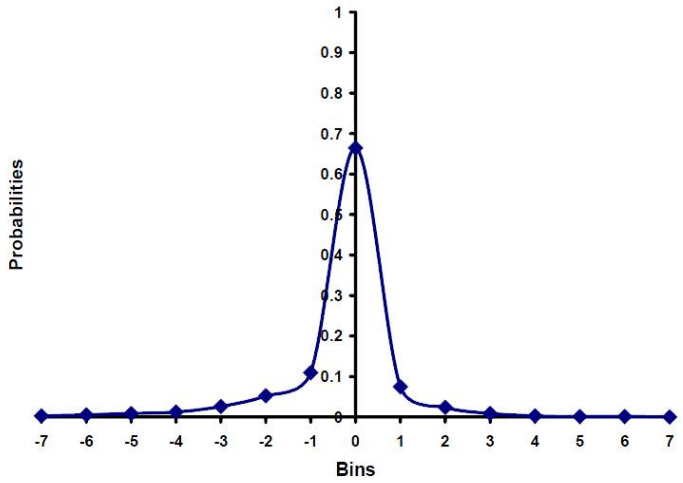

Fig. 11 The average distribution for all customers

to be the most common behavior (67%). Under-estimation (indexed from -1 to -7) is a more common behavior (21%) compared to over-estimation (indexed from 1 to 7). The main reason for customers to under-estimate is to keep a good reputation with the supplier. In our problem environment, capacity is not a main concern for the supplier. The parts are highly customized and excess production is very costly for the supplier. Therefore, the cost of under-estimation is less compared to the cost of over-estimation for the supplier. In a long-term CPFR implementation, this leads to the manufacturers provide forecasts which are with high confidence going to be used. In this way excess production is minimized and the manufacturer builds a reputation with the supplier for accurate forecasts.

6 Managerial Use

Our customer forecast behavior analysis consists of many steps as discussed in the previous sections. Our procedure can be generalized as follows:

1. Orders are transformed into daily flows.
2. Accuracy ratios for daily forecasts are obtained by using the daily flows.
3. A quantitative representation of each customer's forecast is computed by customer bin analysis.
4. Customer order distributions are used as inputs to clustering.
5. The clusters are projected on a two-dimensional graph by using Sammon's Mapping in order to visualize the clusters.

Our customer forecast behavior analysis has several useful applications for the practitioners.

Automation of Detecting Irregular Customer Behavior: When the data set is too large, it is hard to recognize irregularities in a particular customer's forecast. Signalling mechanisms can be equipped to give quick responses. For example, a group of customers can start to give aggressive orders, which directly affect the capacity decisions of the supplier. An automated signaling mechanism can be designed with the help of K-Medoid Analysis. K-Medoid Analysis is based on assigning each customer to the closest cluster center. Therefore, when a customer behaves differently from the other customers, he either initiates a new cluster or joins the cluster of other customers with similar behaviors. Therefore, this gives a signal to the supplier to take necessary actions.

Creating Performance Benchmarks: The forecasting performance of a customer is dependent upon many outside factors such as trends in the industry, competition, exchange rates and etc. It is hard to quantify the effect of each factor on a customer's forecast performance. Therefore, the supplier needs to design flexible performance benchmarks which quickly and dynamically adapt to changes without any supervision. Our clustering analysis discussed in Section 5.2 forms customer groups with similar forecast performances. The cluster with high forecast accuracy represents the ideal customer behavior. As discussed in Section 5.2.2, Kohonen Network is a clustering technique which provides a good spread of customers to the clusters. By performing Kohonen Analysis for every quarter, the supplier can dynamically update the properties of an ideal customer and report this to the customers as performance benchmarks. Kohonen clusters are topological maps and the distance between the Kohonen clusters represents the level of dissimilarity. In Figure (8), from cluster 1 to cluster 8, the percentage of the customers who provides true forecasts decrease. Therefore, as the cluster number increases, the magnitude of the deviation from the ideal behavior increases.

Group Monitoring Options: A customer's forecast behavior can be affected by other customers' actions. Therefore, understanding behavior of related customers is important in decision-making. Customer groups can be formed for close monitoring. For example, in the introduction of a new item to the market, customers are in the initial phases of learning the market conditions. A separate analysis can be performed on this group of customers who are ordering the item. For the items with limited supplier capacity, we can observe aggressive ordering from the customers in order to get more of the supplier's capacity.

Quarterly Reports: Quarterly reports can be generated by the supplier based upon the forecast performance of the customers. Customers can be informed about their performance and make improvements in their forecasting process. Our mapping analysis can be used at that point in order to provide the position of the customer with respect to the other customers. Reward/penalty schemes can be developed based on the performance reports of the customer. The trajectory of a customer for different quarters can be interpreted with respect to different points on the mapping (such as the position of ideal cus-

tomer, his previous quarter position or other customers' or competitors' performances).

Customer-Level Analysis: Our analysis can be repeated for different parts of the same supplier. By doing that, the supplier can understand if the customer is consistent with his behavior for all the parts he is ordering. Parts can be an effective factor on the performance of a customer. A customer can be an over-estimator for one part. However, the same customer can turn out to be a true-estimator for other parts.

7 Conclusion and Extensions

In our analysis we describe complexities such as order-splitting, order-combining and shifts in due dates. We disaggregate the orders into daily flow analysis to overcome complexities and to compute the accuracy ratios for the forecasts. Another solution for handling the complexities is aggregating the orders. In our recent analysis with year 2002 and 2003 data, we search for the minimum aggregation level that eliminates the complexities with the orders. The major disadvantage of the aggregation approach is the loss if information. However, aggregation transforms the data into a format that is compatible with standard statistical parametric techniques.

We employ customer bin analysis to obtain customer order distributions. This vector has been used in our analysis as a quantitative representation of a customer's performance. Using the vectors obtained from customer bin analysis, we categorize customer order distributions into data clusters and use projection techniques for visualization. We derive a distance metric compatible with our problem environment in order to compute the similarities among the customers.

In our analysis we show that customers are consistent with their forecast behavior. Some customers consistently provide bad forecast performance. The supplier should take the necessary actions to improve the forecasts of these customers. However, there are several supply chain features that complicate handling customer forecasts. Forecast effort is a hidden action of the customer. It may be costly for the customer to spend time and effort to provide accurate forecasts. Decreasing forecast effort increases the forecast variability. Since the forecast action is not verifiable, it is also not contractable. The supplier can overcome problems with the forecast variability by rewarding customers based upon observable outcomes (their position in clustering and mapping).

When a customer has a conflict of interest with the supplier, he can provide biased forecasts by inflating or deflating his order size. Over-estimation causes excess production for the supplier and under-estimation leads to lost sales. When the sum of customer orders exceeds the suppliers' capacity, capacity rationing mechanisms lead customers to over-estimate the orders. However, we observe that under-estimation is a more common behavior in our analysis. Therefore, capacity is not the main concern of the customers while placing orders. Reputation concerns become more critical and cause customers to

under-estimate. Since the relationships are long term in the automotive industry, customers (auto manufacturers) tend to keep good relationships with the supplier. The analysis of dynamic models with reputation effect can provide insights into eliminate the bias in customers' forecasts.

References

[Armstrong(2001)] Armstrong JS (2001) Extrapolation for time-series and cross-sectional data. In: Armstrong JS (ed) Principles of Forecasting:A Handbook for Researchers and Practitioners, Kluwer Academic Publishers

[Aviv(2001)] Aviv Y (2001) The effect of collaborative forecasting on supply chain performance. Management Science 47:1326–1343

[Aviv(2002)] Aviv Y (2002) Gaining benefits from joint forecasting and replenishment processes: The case of auto-correlated demand. Manufacturing and Service Operations Management 4:55–74

[Aviv(2007)] Aviv Y (2007) On the benefits of collaborative forecasting partnerships between retailers and manufacturers. Management Science 53(5):777–794

[Azoury(1985)] Azoury KS (1985) Bayes solution to dynamic inventory models under unknown demand distribution. Management Science 31:1150–1160

[Cachon and Lariviere(1999a)] Cachon GP, Lariviere M (1999a) Capacity choice and allocation: Strategic behavior and supply chain performance. Management Science 45:1091–1108

[Cachon and Lariviere(1999b)] Cachon GP, Lariviere M (1999b) An equilibrium analysis of linear, proportional and uniform allocation of scarce capacity. IIE Transactions 45:835–849

[Chen et al(2000)Chen, Ryan, and Simchi-Levi] Chen F, Ryan JK, Simchi-Levi D (2000) The impact of exponential smoothing forecasts on the bullwhip effect. Naval Research Logistics 47:269–286

[Graves(1986a)] Graves SC (1986a) A tactical planning model for a job shop. Operations Research 34:522–533

[Graves et al(1986b)Graves, Meal, Dasu, and Qiu] Graves SC, Meal H, Dasu S, Qiu Y (1986b) Two stage production planning in a dynamic environment. In: Axsater S, Schneeweiss C, ESilver (eds) Multi-Stage Production Planning and Control: Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, pp 9–43

[Graves et al(1998)Graves, Kletter, and Hetzel] Graves SC, Kletter DB, Hetzel WB (1998) A dynamic model for requirements planning with application to supply chain optimization. Operations Research 46:S35–S49

[Gullu(1986)] Gullu R (1986) On the value of information in dynamic production/inventory problems under forecast evolution. Naval Research Logistics 43:289–303

[Hausman(1969)] Hausman WH (1969) Sequential decision problems: A model to exploit existing forecasters. Management Science 16:B93–B111

[Heath and Jackson(1994)] Heath DC, Jackson P (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. IIE Transactions 26:17–30

[Kohonen(2001)] Kohonen T (2001) Self Organizing Maps. Springer

[Lariviere and Porteus(1999)] Lariviere MA, Porteus EL (1999) Stalking information: Bayesian inventory management with unobserved lost sales. Management Science 45:346–363

[OICA(2009)] OICA (2009) 2009 production statistics. Tech. rep., International Organization of Motor Vehicle Manufacturers

[Sammon(1969)] Sammon JW (1969) A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C-18:401–409

[Scarf(1959)] Scarf H (1959) Bayes solutions of the statistical inventory problem. Annals of Mathematical Statistics 30:490–508

[Taylor and Xiao(2010)] Taylor TA, Xiao W (2010) Does a manufacturer benefit from selling to a better-forecasting retailer? Management Science 56(9):1584–1598

[Terwiesch et al(2003)Terwiesch, Ren, Ho, and Cohen] Terwiesch C, Ren JZ, Ho TH, Cohen M (2003) An empirical analysis of forecast sharing in the semiconductor equipment industry. Working paper, Wharton Business School, University of Pennsylvania, Philadelphia, PA

[Tischendorf et al(2008)Tischendorf, Handschuh, Landgraf, and Romero] Tischendorf J, Handschuh M, Landgraf A, Romero R (2008) Automotive suppliers: How to grow organically and sustainably. Tech. rep., A.T. Kearney

[Toktay and Wein(2001)] Toktay B, Wein LM (2001) Analysis of a forecasting-production-inventory system with stationary demand. Management Science 47:1268–1281