Stochastics and Statistics

Forecasting demand for single-period products: A case study in the apparel industry

Julien Mostard [a], Ruud Teunter [b,*], René de Koster [a]

[a] RSM Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[b] University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

ABSTRACT

The problem considered is that of forecasting demand for single-period products before the period starts. We study this problem for the case of a mail order apparel company that needs to order its products pre-season. The lack of historical demand data implies that other sources of data are needed. Advance order data can be obtained by allowing a selected group of customers to pre-order at a discount from a preview catalogue. Judgments can be obtained from purchase managers or other company experts. In this paper, we compare several existing and new forecasting methods for both sources of data. The methods are generic and can be used in any single-period problem in the apparel or fashion industries. Among the pre-order based methods, a novel 'top-flop' approach provides promising results. For a small group of products from the case company, expert judgment methods perform better than the methods based on advance demand information. The comparative results are obviously restricted to the specific case study, and additional testing is required to determine whether they are valid in general.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In the apparel industry, three prominent developments contribute to the complexity of forecasting: shortening product life-cycles, increasing product variety, and globalization of sourcing and manufacturing. The impact of each of these developments will be discussed in detail.

Ninety-five percent of SKUs (Stock Keeping Units) change every selling season (Gutgeld and Beyer, 1995). Because product life-cycles are short, there are no historical demand data that can be used to obtain a priori demand forecasts, i.e. before any demand has been realized. Furthermore, the number of in-season replenishment opportunities after observing demand and updating forecasts accordingly is limited and the risk of product obsolescence is high.

Due to global competition, faster product development, technological advances, increasingly flexible manufacturing systems, and more demanding consumers, an unprecedented number and variety of products are competing for demand (Fisher et al., 1994). As a result, the volume of sales per SKU is very low (Gutgeld and Beyer, 1995), and demand for SKUs within the same product line can vary significantly (Abernathy et al., 2000). Thus, even if aggregate demand can be predicted with some certainty, it is very difficult to predict how that demand will be distributed over the many products that are offered. The complexity of production planning and ordering increases accordingly.

The bulk of products is produced in South East Asia, and hence the lead time to Western retailers is long. The typical lead time from fabric manufacturers is 3 months (Gutgeld and Beyer, 1995). The specific mail order company that we study faces lead times of 6–14 weeks. Long lead times dictated by powerful suppliers and strong competition from other retailers trying to secure enough production capacity force retailers to commit to initial order quantities long, usually several months, before the start of the selling season. We remark that there are a few well-known apparel/fashion retailers like Zara and Hennes & Mauritz that have deviating strategies based on local manufacturing. However, for the large majority of apparel retailers, who have a low-cost focus and often sell private label products, the situation is as we described.

So, for most products that an apparel retailer sells in any season, a demand forecast is needed well before the start of the season, when no historic demand data is available. An apparel retailer's success hinges to a large extent on the accuracy of those pre-season forecasts on which the initial orders are based. These initial orders comprise the bulk of the total volume ordered (Fisher and Raman, 1999). Additional in-season replenishment opportunities, if available, are essentially emergency replenishment opportunities and only serve to prevent shortages resulting from possible initial underestimation of demand. We refer interested readers to Mostard and Teunter (2006) for a further discussion and analysis of inventory control issues. In this paper, the focus is on forecasting.

To obtain maximum accuracy of initial forecasts, there are two common practices for gathering relevant information. First, a so-called *preview* is common in the apparel industry (see, e.g., Tang

* Corresponding author. Tel.: +31 50 3638617.
 E-mail address: r.h.teunter@rug.nl (R. Teunter).

et al., 2004; Fisher and Rajaram, 2000; Chambers and Eglese, 1986). During a two to five week period before the start of the selling season (the *preview period*), customers can pre-order products at a small discount. Second, many apparel retailers use a committee of experts (e.g. purchasers, planners) to provide forecasts for individual products (see, e.g., Mantrala and Rao, 2001; Fisher et al., 2000; Raman, 1999).

In this paper, we propose new forecasting methods based on advance demand information, and perform a case study to compare them to existing ones based on advance demand information and also to methods based on expert judgments. Numerical results are obtained using data from a large mail order/Internet retailer based in the Netherlands. This company currently bases its forecasts on advance demand information. Based on a data set of around seven hundred SKUs and for two successive summer seasons, we compare the accuracy of the various methods based on advance demand information. For a smaller subset of around one hundred SKUs, we also obtained forecasts from a number of company experts. For this subset, we compare methods based on these expert judgments to methods based on advance demand information.

The remainder of the paper is organized as follows: We will first give an overview of the relevant literature in Section 2 and outline our contributions. In Section 3, we describe the different forecasting methods. Section 4 introduces the case company and available data in more detail. The empirical results are described in Section 5. Finally, we present our conclusions, discuss limitations, and provide directions for further research in Section 6.

2. Literature review: forecasting demand for single period (fashion) products

Fashion products in general are characterized by high demand uncertainty, high stockout costs and a high risk of obsolescence (Lee, 2002). Although the specific mail order company that we study can be classified as an apparel company rather than a fashion company, it shares these characteristics. This is evidenced by the fact that the company frequently has significant leftovers of individual SKUs which cannot be carried over to the next season and need to be sold at high markdowns. Customer satisfaction and retention are crucial in the mail order business, and the company can therefore not afford to run out of stock on many SKUs, as that would turn away customers. Therefore, we review the literature on apparel as well as fashion companies, and more generally on single period/single season products.

Raman (1999) finds that few fashion companies are aware of, let alone use, the mathematical models for fashion planning that have been proposed in the literature. He notes that most papers fail to demonstrate the proposed models using applications and to provide thorough evidence of their ability to influence managerial decisions. Other important shortcomings are that most proposed methods rely on demand data gathered using the selling season (a posteriori forecasting), and do not consider expert judgment.

In the remainder of this section, we first mention some papers that do not deal with forecasting but related management problems, then shortly discuss a posteriori forecasting, and finally discuss the a priori methods in detail as they are the most relevant for our study. The a priori methods use either historical data or expert judgment and the relevant contributions are discussed in Sections 2.1 and 2.2, respectively.

A number of authors (Gallego and Ozer (2001), Mostard et al. (2005), Kogan et al. (2008), Kogan and Herbon (2008), Wang and Toktay (2008)) discuss the complications of limited demand information and forecasting inaccuracy for new SKUs on production and inventory planning for those SKUs.

Contributions on a posteriori forecasting using in-season demand include Hertz and Schaffir (1960), Murray and Silver (1966), Chang and Fyffe (1971), Green and Harrison (1973), Chambers and Eglese (1988), Fisher and Raman (1996), Hill (1997), Sethi et al. (2003), Yan et al. (2003), Choi (2007), Sethi et al. (2007), Rahman (2008), Au et al. (2008). Fisher and Raman (1996) show using an example that using in-season demand, if only for a few weeks, can considerably improve forecasts. Their advice to retailers is therefore to first observe this initial demand and then order. However, as we have argued in the previous section, retailers often do not have this option in practice, and we will therefore concentrate on a priori forecasting.

2.1. A priori forecasting using historical data

Chang and Fyffe (1971) assume that a firm has a "long-run sales history of individual seasonal-style-goods SKUs or groups of such SKUs". They propose to estimate demand by using regression on those historic sales, also based on the "outcome of some observable variable". However, they do not explain in detail how that can be done nor do they test the method using real data. It seems difficult to apply this method in the apparel industry, as long-run sales histories of very similar products are rare.

Chambers and Eglese (1988) discuss the use of preview demand data that are gathered by sending out a preview catalogue (which does not necessarily include a full product range) to a sample comprised of several thousand regular customers and offering them the opportunity to order products at a discount before the season starts. They assume that an aggregate forecast for the full product range is given, and propose to forecast the demand for a product line by multiplying the aggregate forecast with the fraction of total preview demand for products in that product line. They further propose a second, slightly more sophisticated forecasting method, which takes into account that the ratio of total demand to preview demand ('the scaling factor') may not be the same for all product lines. These methods are very suitable and, indeed, have been developed for an apparel mail order company.

Thomassey and Happiette (2007) propose a decision-support system based on neural networks, which automatically performs item sales forecasting. The system is designed to deal with many characteristics of the apparel market: large number of items, short lifetimes, substitution of most items with each new collection, long lead times, and influence of many external factors like the weather, promotions, fashion, and the economic environment. The proposed system is composed of three steps: obtain prototypes of demand behavior using a clustering procedure on historical demand data, (2) link these prototypes to descriptive criteria (e.g. price, lifespan or materials) using a probabilistic neural network, and (3) assign each new item to a prototype based on the item's descriptive criteria. Forecasts generated by the proposed model on a set of 285 new items from a French apparel distributor have a MAPE of 147%. So, accuracy is low despite the complexity of the method. For this reason, we decided not to include this method in our comparative study. The results in Section 5 will show that the simpler methods that we do consider are all more accurate (for our data set).

2.2. A priori forecasting based on expert judgment

Green and Harrison (1973) discuss the use of product comparisons by a consumer panel (female members of the company) for estimating demand. They propose a rather complex forecasting method. Demand is modeled as being log-linearly related to the number of votes by the panel and the price, and the parameters of that relation are estimated in a complicated way using sales data from previous seasons. This complexity reduces the applicability of

the method for the situation of a mail order apparel retailer that we consider, although the situation that they consider is similar.

Fisher and Raman (1996), Fisher et al. (2001) propose to let a number of experts within a company estimate the demand for a product. The demand is calculated as the average of the experts' estimates. The method is straightforward and very applicable for the specific company that we consider. Indeed, both Fisher and Raman (1996), Fisher et al. (2001) also proposed the method for a retailer in the apparel industry.

Mantrala and Rao (2001) also develop two forecasting methods based on experts' estimates for an apparel retailer (of the demand for mens walking shorts for the spring season). Their methods are more detailed than those of Fisher and Raman (1996), Fisher et al. (2001), since they divide the season into a number of periods and also consider different price levels. The first method starts by asking each expert separately for the minimum, maximum, and most likely (modus) demand for each combination of period and price. Subsequently, using the Delphi group method, the experts have to reach consensus on the minimum, maximum, and modus for each combination of period and price. Finally, for each combination of period and price, the forecast (for the mean) is calculated as the average of the minimum, the maximum, and the modus. The second method asks different input from the experts: an estimate of total (over all periods) demand at a single price, as well as a 95% confidence interval; the expected percentage of total demand that will occur in each period; and an estimate of the price elasticity. Based on these inputs, total demand is estimated using a rather complicated model including a log-normal disturbance term. The authors do not report any results on the quality of the resulting forecasts of the two methods.

Based on a survey among 240 firms, Sanders and Manrodt (2003) report judgmental forecasting methods to perform less well than quantitative methods. They offer two explanations for the poor performance of judgmental forecasting. First, there are a number of inherent biases, including optimism, wishful thinking, lack of consistency, political manipulation, and overreacting to randomness. Second, people have a limited ability to consider and process large amounts of information.

On the other hand, judgmental methods are often preferred by practitioners, since they can incorporate special insights, trends, and macro-economic factors, which are hard, if not impossible, to quantify in practice and since practitioners are more acquainted with them. Moreover, a lack of data often rules out the use of complex forecasting methods. This is certainly true for the mail order retailer that we consider.

3. Forecasting methods

We will discuss methods that forecast based on advance demand analysis in Section 3.1, and then continue in Section 3.2 with expert judgment methods. For all methods, we let $\mathcal{N}$ denote the set of SKUs in an upcoming selling season for which demand forecasts are needed, and N denote the number of SKUs in $\mathcal{N}$.

3.1. Methods based on advance demand information

We consider three methods based on advance demand information. Each of these methods first forecasts total season demand in the upcoming season, denoted by M, for a group of SKUs $\mathcal{N}$ by scaling up the registered advance (preview) demands for those SKUs, and then divides this forecasted group demand over the individual SKUs. The scaling up factor is calculated as the ratio of final demand to preview demand for a 'comparable' group of SKUs (e.g. t-shirts), denoted by $\mathcal{H}$, in one or more historical seasons.

Using notation P_n for the preview demand in the new season for SKU $n \in \mathcal{N}$, H for the number of SKUs in $\mathcal{H}$, R for the preview demand in the historic season(s) for SKU $h \in \mathcal{H}$, and S_h for the total demand in the historic season(s) for SKU $h \in \mathcal{H}$, this gives

$$M = \frac{\sum_{h \in \mathcal{H}} S_h}{\sum_{h \in \mathcal{H}} R_h} \sum_{n \in \mathcal{N}} P_n. \tag{1}$$

So given preview demand, we forecast total demand by assuming that the ratio of total demand to preview demand will be the same as in past season(s) for a comparable group of SKUs. We remark that this forecast could be modified if additional information on e.g. the economical situation or meteorological conditions were available.

All methods can be applied for any choice of grouping. Intuitively, it makes sense to group SKUs in such a way that the SKUs in $\mathcal{N}$ have similar product characteristics as the SKUs in $\mathcal{H}$. Note that in order to obtain a decent estimate of M, it is required that SKUs in historical season(s) can be found that bear sufficient resemblance to the SKUs in $\mathcal{N}$. In our numerical investigation, we will consider several ways of grouping in line with classifications used by the case company (see Section 4).

Collections change every selling season to follow the latest fashion and trends. Hence, there will generally be no overlap between $\mathcal{H}$ and $\mathcal{N}$, although there might be a group of generic SKUs that are carried over from one season to the next. While individual SKUs change, the definitions of groups and the classification of SKUs into these groups do not change.

Note that in the unlikely event that $\mathcal{H} = \mathcal{N}$, (1) would result after rearranging terms in $M = \frac{\sum_{n \in \mathcal{N}} P_n}{\sum_{n \in \mathcal{N}} R_n} \sum_{n \in \mathcal{N}} S_n$. So, the relative increase in total demand is forecasted to be equal to the relative increase in the preview demand, which is logical.

The forecast m_n for SKU n, $n = 1, \ldots, N$, is obtained by taking fraction f_n of M, i.e.

$$m_n = M f_n, \quad n \in \mathcal{N}. \tag{2}$$

The three methods that use advance demand information differ in the calculation of the fractions f_n. The fractions are proportional to preview demand for Method 1, equal for Method 2, and based on a 'top-flop' division for Method 3. Details will be provided below for all methods. We remark that Method 1 was first proposed by Chambers and Eglese (1988). To the best of our knowledge, the other division methods are new to the literature.

3.1.1. Method 1 (preview division)
Preview division divides M proportional to preview demand, i.e., each SKU $n \in \mathcal{N}$ gets fraction

$$f_n = \frac{P_n}{\sum_{n \in \mathcal{N}} P_n}$$

of M. Using (1) and (2), we get that

$$m_n = \frac{\sum_{h \in \mathcal{H}} S_h}{\sum_{h \in \mathcal{H}} R_h} P_n, \quad n \in \mathcal{N}.$$

This method is included because it is used by the case company, in combination with Method 2. Preview division was previously proposed by Chambers and Eglese (1988). However, they restrict the method to grouping SKUs per 'product line', while here it can be applied for any chosen way of grouping.

3.1.2. Method 2 (equal division)
Equal division divides M equally over the SKUs in $\mathcal{N}$. Thus,

$$f_n = \frac{1}{N}.$$

There are two main reasons for including this simplistic approach. First, the approach is used by the case company, in combination with Method 1. Second, this approach serves as a benchmark for interpreting the performance of other methods.

3.1.3. Method 3 (top-flop division)

Top-flop division is based on the idea that the demand percentages of the 'top' and the 'flop' SKUs in a group of SKUs are fairly stable over time. For example, the 33% best-selling SKUs in a product group of t-shirts represent about 70% of total t-shirt demand, while the 33% worst-selling t-shirts only represent 5% (and the remaining 33% of SKUs represent 25%). Note that the number of top-to-flop categories is 3 in this example, but can be any number in general. We remark that this bears similarities with the classification of SKUs in ABC systems, which are common in practice. The idea has also been suggested by experts from the case study company. This method is included because it is new and intuitively attractive.

So, the historical SKUs in $\mathcal{H}$ are divided into a set C equally sized categories, ranging from top (highest total demand) to flop (lowest total demand). These categories are numbered $1, \ldots, C$ and the corresponding subsets of $\mathcal{H}$ are denoted by $\mathcal{H}_1, \ldots, \mathcal{H}_C$. If H/C is integer, then each category contains H/C SKUs. Otherwise, categories $1, \ldots, H - \lfloor H/C \rfloor C$ contain $\lceil H/C \rceil$ SKUs and the remaining categories contain $\lfloor H/C \rfloor$ SKUs. Here, $\lfloor x \rfloor$ denotes x rounded down to the nearest integer and $\lceil x \rceil$ denotes x rounded up to the nearest integer. Let H_c, $c = 1, \ldots, C$, denote the number of SKUs in $\mathcal{H}_c$.

Similarly, the SKUs of $\mathcal{N}$ are divided from top to flop into C categories denoted by $\mathcal{N}_1, \ldots, \mathcal{N}_C$, but based on preview demand. Let N_c, $c = 1, \ldots, C$, denote the number of SKUs in $\mathcal{N}_c$ and let $c(n)$, $n \in \mathcal{N}$ denote the category in which SKU n falls. In case N/C is non-integer, SKUs are divided over the categories in the same way as described above for $\mathcal{H}$.

The fraction of historic demand in each category c, corrected for the number of SKUs per category, is

$$g_c := \frac{\frac{1}{H_c} \sum_{h \in \mathcal{H}_c} S_h}{\sum_{c=1}^{C} \frac{1}{H_c} \sum_{h \in \mathcal{H}_c} S_h}, \quad c \in \mathcal{C}.$$

Note that the fractions over all categories sum to 1.

Each SKU $n \in \mathcal{N}$ receives fraction

$$f_n = \frac{g_{c(n)}}{\sum_{n \in \mathcal{N}} g_{c(n)}} = \frac{g_{c(n)}}{\sum_{c=1}^{C} N_c g_c}, \quad n \in \mathcal{N}.$$

of the estimated demand M. Note that these fractions sum to 1.

For the special case that N/C and H/C are integer (so that all categories contain equally many SKUs), the above formulae simplify to

$$g_c = \frac{\sum_{h \in \mathcal{H}_c} S_h}{\sum_{h \in \mathcal{H}} S_h}, \quad c \in \mathcal{C}$$

and

$$f_n = \frac{g_{c(n)}}{N/C}, \quad n \in \mathcal{N}.$$

3.2. Expert methods: general description

We describe (variants of) two expert judgment methods (numbered 4–5) that have been suggested in the literature for pre-season forecasting of demand in the apparel industry (see also Section 2.2).

3.2.1. Method 4 (experts' average)

This method simply calculates the average of a number of expert estimates. Let E denote the number of experts, and $m_{n,e}$ denote the forecast of expert e, $e = 1, \ldots, E$, for SKU $n \in \mathcal{N}$. Then the average experts' forecast m_n for SKU n can be written as

$$m_n = \frac{1}{E} \sum_{e=1}^{E} m_{n,e}, \quad n \in \mathcal{N}.$$

This averaging method was applied by Fisher and Raman (1996), Fisher et al. (2000), who use panels of 7 and 4 experts, respectively. The use of an expert panel (also referred to as a buying committee) in forecasting has also been documented by Raman (1999), Mantrala and Rao (2001). Combining forecasts of experts increases accuracy, because inconsistencies of one expert tend to cancel out the inconsistencies of another (Blattberg and Hoch, 1990).

3.2.2. Method 5 (expert triangulation)

Method 5 is the first of two methods proposed by Mantrala and Rao (2001) and has been reviewed in Section 2. We use a simplified version, with fixed prices and for a single period. Furthermore, instead of asking the experts to reach consensus on the minimum, maximum and modus of the demand and then taking the average of these three consensus figures as the forecast, we calculate the averages over the minima, maxima and modi from all experts (thereby weighing their forecasts equally).

This method is included for two reasons. First, inclusion of this method in our study allows for a comparison with the very simple expert method (Method 4), in order to determine whether the slight additional sophistication introduced in this method leads to better forecasts. Second, it has not been tested (by the authors who proposed it).

4. Application: a large mail order/Internet retailer

The case company is a mail order/Internet apparel retailer operating only in the Netherlands. It divides each year into two selling seasons, spring–summer (December–June) and autumn–winter (June–December). One main catalogue is issued per season, and several smaller catalogs appear throughout the year, containing special collections or special offers aimed at specific groups of customers. A total of around 80,000 SKUs are offered, distributed over three collections: apparel, furniture and consumer electronics. We focus on the apparel collection in the main catalog, which generally contains around 25,000 SKUs.

Per sales season, the company distributes over 70,000 preview catalogs and 1.5 million main catalogs, generating roughly 6 million order lines. The company's yearly turnover is around 350 million Euro. Apart from being the largest home shopping retailer in the Dutch market with 1.5 million returning customers, the company has also become a major and successful Internet retailer over the past few years. It uses four sales channels: website, call center, voice response system, and regular mail. Over the three years covered by our data set, the company's turnover has decreased from 357 million euros, to 339 million, and 326 million due to consumers becoming more price-sensitive. In the same period, orders via the website have increased from 20% of the total volume, to 30%, and over 40%.

Based on product characteristics, SKUs are categorized into assortment groups, which are subdivided into product groups. Thus, SKUs in a product group share more characteristics than SKUs in an assortment group. While individual SKUs are different from one season to the next, the categorization of SKUs into product and assortment groups does not change. For example, a product group may consist of all ladies' singlets sold in a particular season, which are a subset of the assortment group with ladies' upper wear.

4.1. Data

The company has provided data of a subset of private label apparel products that appeared in the main catalogs of three successive spring–summer seasons, which we will refer to as Seasons 1, 2 and 3. The data are taken from three assortment groups that all belong to 'ladies actuals'. The company has labeled these groups as: outer wear, tops, and bottoms. We will refer to them as Assortment groups 1, 2 and 3, respectively.

For all three seasons and all SKUs in these assortment groups, the preview demand and the total sales are given. Using additional data on lost sales, the total demand can be estimated, as will be discussed next. After doing so, we will describe how expert judgment data was collected for a subset of SKUs in Season 3.

4.2. Sales and lost demand data

The total sales are not necessarily equal to total demand, since some demand may have been lost. For the case that lost demand is not recorded at all, Fisher et al. (2000) propose to estimate lost demand based on the moment at which the stock dropped to zero and the demand curve until that time. However, that method cannot be applied in our case, since the moment at which the stock drops to zero is not recorded either. Fortunately, lost demand is partially recorded in our case, since the company does register lost demands that are received through the call center and the voice response system (but not through the website and regular mail). Moreover, we know the percentage distribution of total demand over the four sales channels. This information is used to scale up the registered lost demand for each SKU to get our estimate of total lost demand per SKU.

We refer interested readers to Bell (2000) for a discussion of estimating the distribution of demand based on sales data for multiple-period problems.

4.3. Expert judgment data

We used a panel of seven experts (i.e. $E = 7$), consisting of three supply chain controllers, a category buyer, a supply chain manager, a commercial assistant, and a buying manager. They had been with the company for an average of about 5 years. Each expert was asked to give independent estimates of total demand and its range (i.e. the minimum and maximum) for all 89 SKUs in Assortment group 1 of Season 3. This Assortment group was chosen as it contained the smallest number of SKUs of the three groups considered. This limited the time involved for the group of experts, and thereby ensured that they stayed concentrated throughout the experiment. Based on the similarities in the assortment groups (all belonging to ladies' actuals), similar results were expected for all three.

The experts are all involved in the forecasting and ordering processes of these products on a daily basis. They were not given a budget restriction. The experts' judgments were made before preview demand information was available.

4.4. Initial data exploration

Some aggregate figures concerning the available data are shown in Table 1. The sizes of both the assortment groups and the product groups vary greatly across the groups. In Season 3, the numbers of SKUs in assortment and product groups ranged from 89 to 322 and from 4 to 195, respectively.

Note from Table 1 that the ratio of realized demand to observed preview demand is reasonably constant over the three seasons. This is important, since Methods 1–3 all use this ratio to forecast at the aggregate level.

Table 1
Aggregate data from Seasons 1–3.

	Season 1	Season 2	Season 3
Number of SKUs	801	700	628
Total preview demand (units)	2669	2443	2497
Total realized sales (units)	398,404	403,483	313,985
Total registered lost demand (units)	11,206	35,148	38,921
Total estimated lost demand (units)	16,013	58,593	74,859
Average ratio of realized to preview demand	200	223	217

It also appears from Table 1 that the total registered lost demand increases over time, especially from Season 1 to Season 2. A possible explanation given by the company is that registration of lost demand has improved over the years. However, no estimates on the effect of improved registration were provided. Therefore, and also after checking that more stable lost demand figures did not lead to substantially different results and conclusions, the registered lost demand figures were left unchanged.

Next, we explore the correlation of preview demand and expert judgment with the realized sales. This is done graphically in Figs. 1 and 2.

As expected, both figures indicate a positive correlation. The Pearson correlation coefficients are 0.54 and 0.74 for Figs. 1 and 2, respectively, and the corresponding P values are less than 0.001 for both cases, indicating that there is positive correlation even at the 0.1% significance level. However, it is also apparent that
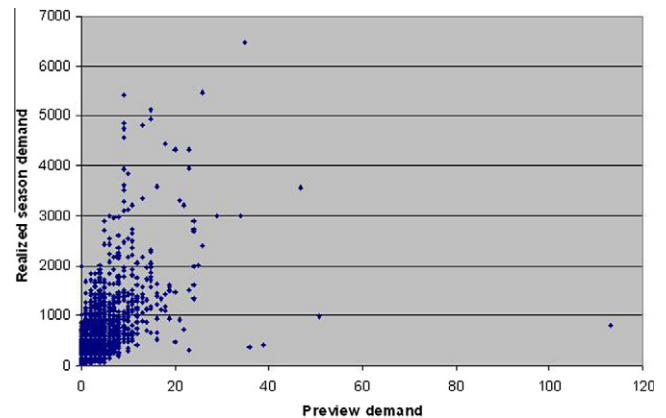

Fig. 1. Realized demand against preview demand (Seasons 1–3; Assortment groups 1–3).
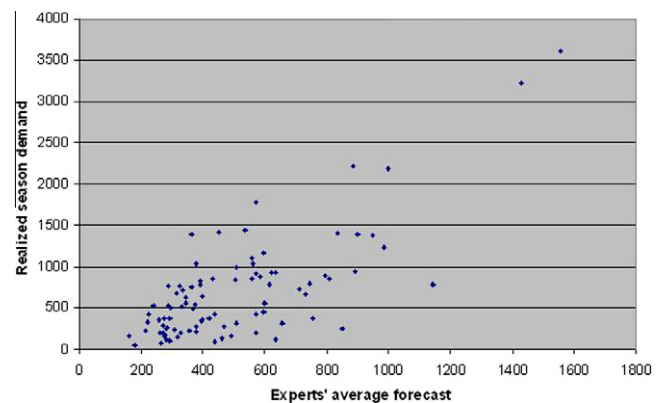

Fig. 2. Realized demand against the average experts' estimate (Season 3, Assortment group 1).

there are many outliers, and hence we can expect reasonably large forecast errors for all forecasting methods.

4.5. Methods: case specific settings

The case company combines SKUs into product groups and product groups into assortment groups. The methods based on advance demand information (Methods 1–3) can therefore be on a product group level, on an assortment level, or for all assortment groups together. We tested all three ways for all three methods, and for each method it turned out that forecasting on a product group level provided the best results (and for all SKUs together the worst). Therefore, we will only report the results on a product group level.

We remark that the ongoing policy of the company was actually to produce forecasts based on advance demand information, but on an assortment group level. Their method used a combination of the rules for dividing group demand (i.e. for calculating f_n, $n \in \mathcal{N}$) of Methods 1 and 2. It applied a different way of estimating group demand M. Instead of using (1), a planning committee consisting of mainly purchasers had to reach a consensus, also taking budget restrictions into account. Since the resulting forecasts have not been recorded, they cannot be compared to those of other methods in our empirical investigation. We do remark that letting budget restrictions play a role in forecasting obviously carries the risk of underestimations to stay within budget or over-estimations to avoid losing part of the budget (in future years).

Method 3 (top-flop) was tested with 3–7 categories, and it was found that the number of categories had negligible effect on the results. Since having a top, mid and flop class is intuitively most appealing, only results for three categories are presented. We also tested the top-flop method with varying class sizes. However, again, this did not (significantly) improve the performance. Therefore, we report results for equal-sized classes only. We note that contrary to ABC inventory classification, where class A SKUs typically get special attention and their number therefore needs to be limited, class sizes do not affect the complexity of applying the top-flop method.

5. Empirical results

In this section, we will compare the five forecasting methods using the case study data described in Section 4. Methods 1–3 will first be compared for the full data set (assortment groups 1–3) and in their forecasting accuracy for Season 2 (based on Season 1) and Season 3 (based on Season 2). Then, for Assortment group 1 and Season 3, Methods 1–3 will also be compared to Methods 4 and 5 based on expert judgment.

We used three different performance measures of forecast accuracy: mean absolute percentage error (MAPE), mean absolute deviation (MAD) and mean percentage error (MPE). The comparative performance of the different methods was consistent over the three error measures, and hence we report the results for MAPE only. We remark that we did not consider the mean square error (MSE) because of its sensitivity to outliers.

We report overall results (per year) as well as for classes of SKUs that are based on the value of the preview demand P. The classes that we use are $P = 0$, $0 < P \leqslant 2$, $2 < P \leqslant 5$, $5 < P \leqslant 10$, $P > 10$. We use this classification, since we expect that higher preview demands will imply more reliable statistical extrapolations of that demand, and will therefore influence the relative performance of the different methods, especially statistical (Methods 1–3) versus expert judgment (Methods 4–5). We remark that the overall results exclude SKUs with $P = 0$, i.e. report the average over all SKUs with positive preview demand. This is done because some of the

methods will always result in a zero forecast for these SKUs, and their inclusion would therefore reduce the meaningfulness of the overall results.

5.1. Methods based on advance demand information

Table 2 gives the average MAPE for all SKUs with positive preview demand together (overall) and also per preview demand class. Furthermore, the error percentages in bold were significantly lower (based on Tukey tests at a 5% significance level) than those for other methods (if any) that are not in bold, but not significantly different from each other.

It appears that all methods perform considerably better in Season 2 than in Season 3. An important contributing factor to the poor performance in Season 3 is that demand dropped sharply compared to previous years, although preview demand was comparable to previous years. This may have been caused by a number of factors, including macro-economic and weather conditions. We discussed this with company experts, but neither they nor we could identify important explanatory market or economic conditions as part of the cause. We remark that all data was collected before the start of the current global recession.

As expected, Method 2 based on equal division performs worst on average. Method 1 (preview) provides the best overall performance. However, as is especially evident for Season 3, Method 1 can lead to large forecast errors for SKUs with high preview demand. Method 1 often results in much too large forecasts for those SKUs. This leads to large stocks remaining at the end of the season that either become obsolete or have to be sold below the cost price. Methods 3 avoids those large forecast errors for SKUs with high preview demand. Apparently, although high preview demand is indeed a reliable indicator of whether an SKU will be top, the exact ranking of the top SKUs based on preview demand is no guarantee that the final ranking based on realized demand will be the same. This is illustrated for a specific product group with 9 SKUs in Table 3. For this group, the three SKUs with the highest (lowest) preview demand indeed turn out to be the top (flop) SKUs. However, the realized demand for the SKU with the highest preview demand of 8, is only about half of that for the SKU with preview demand 7.

For the above example with 9 SKUs, the three top SKUs and the three flop SKUs are all correctly identified. In general, especially for larger numbers of SKUs, the classification is not perfect. However, most SKUs do typically end up in the correct class. To illustrate this, we consider a second example of a product group with 37 SKUs in Season 3. Table 4 shows the preview demand and the

Table 2
Forecast errors (MAPE) averaged over the SKUs of all three assortment groups for Seasons 2 (top) and 3 (bottom). All forecast methods are applied at the product group level. Errors in bold are significantly lower (based on Tukey tests at a 5% significance level) than those (if any) not in bold, but not significantly different from each other.

	Method 1 (preview division) (%)	Method 2 (equal division) (%)	Method 3 (top-flop division) (%)
Season 2			
$P_i = 0$	**100**	263	**67**
$0 < P_i \leqslant 2$	**55**	117	**63**
$2 < P_i \leqslant 5$	**62**	**61**	85
$5 < P_i \leqslant 10$	57	50	53
$P > 10$	69	**54**	**39**
Overall $P > 0$	**59**	83	**67**
Season 3			
$P_i = 0$	**100**	384	**110**
$0 < P_i \leqslant 2$	**51**	158	**71**
$2 < P_i \leqslant 5$	**72**	94	123
$5 < P_i \leqslant 10$	96	**52**	121
$P > 10$	263	**60**	**95**
Overall $P > 0$	**81**	113	**98**

Table 3
Preview demands and realized demands for a specific product group with 9 SKUs in Season 3.

Preview demand	Realized demand
1	291
1	99
1	218
2	316
3	330
3	519
4	534
7	1344
8	684

Table 4
Comparison of advanced demand methods 1 (preview division) and 3 (top-flop division) for a specific product group with 37 SKUs in Season 3.

Preview demand	Realized demand	Method 1 (preview)		Method 3 (top-flop)		
		Forecast	MAPE (%)	Class	Forecast	MAPE (%)
11	1275	4167	227	Top	1372	8
8	1066	3030	184	Top	1372	29
5	673	1894	181	Top	1372	104
4	2006	1515	24	Top	1372	32
4	1867	1515	19	Top	1372	26
4	1833	1515	17	Top	1372	25
4	1040	1515	46	Top	1372	32
4	941	1515	61	Top	1372	46
4	719	1515	111	Top	1372	91
3	716	1136	59	Top	1372	92
3	610	1136	86	Top	1372	125
3	603	1136	88	Top	1372	128
3	509	1136	123	Top	1372	170
3	505	1136	125	Mid	781	55
2	1234	758	39	Mid	781	37
2	749	758	1	Mid	781	4
2	729	758	4	Mid	781	7
2	476	758	59	Mid	781	64
2	471	758	61	Mid	781	66
2	216	758	251	Mid	781	261
2	209	758	262	Mid	781	273
2	181	758	319	Mid	781	331
1	845	379	55	Mid	781	8
1	750	379	49	Mid	781	4
1	635	379	40	Mid	781	23
1	601	379	37	Flop	447	26
1	569	379	33	Flop	447	21
1	558	379	32	Flop	447	20
1	359	379	46	Flop	447	72
0	524	0	100	Flop	447	25
0	420	0	100	Flop	447	15
0	260	0	100	Flop	447	6
0	227	0	100	Flop	447	97
0	204	0	100	Flop	447	119
0	178	0	100	Flop	447	151
0	167	0	100	Flop	447	168
0	152	0	100	Flop	447	194

season demand, as well as the forecasts and associated forecasting accuracies for Methods 1 and 3. For presentational ease, the SKUs are sorted in descending order of preview demand, and SKUs with the same preview demand are sorted in descending order of season demand. The three blocks from top to bottom are classified as top, mid and flop, respectively, by Method 3. We remark that this classification is slightly arbitrary, because SKUs in two classes sometimes have the same preview demand and hence could have been swapped. However, we checked that this does not significantly alter the (average) performance of the 'top-flop' method that will be discussed next. For completeness, we provide the following

relevant data (based on the demand in Season 2) for calculating the forecasts in Season 3: $M = 32,576$, $g_1 = 0.528$, $g_2 = 0.300$ and $g_3 = 0.172$.

Table 4 shows that the classification as top, mid and flop based on preview demand is correct for most SKUs. There are exceptions though. For instance, the fifth best selling SKU with season demand 1234 is not included in the top 13 SKUs based on preview demand. However, 7 of the 8 best selling SKUs are included in the top class.

Note also from Table 4 that none of the three best selling SKUs is included in the top-3 based on preview demand. This explains why for the top SKUs of this product group, in line with the above reported average findings for all SKUs, the top-flop Method 3 is more accurate than Method 1 based on preview division. An additional benefit of Method 3 is that, contrary to Method 1, it does not produce a zero demand forecast for any product (unless zero (preview) demand was registered for all SKUs in the product group). The mail order company will of course never order zero SKUs of any of the products that have been designed for the new season.

It is impossible to correctly identify all top and flop SKUs based on preview demand. Preview demand for most SKUs is low; the vast majority of products are ordered less than 10 times during the preview period. Hence, there are minor or even no differences in preview demand between some of the SKUs that appear in different categories. Yet the differences in total season demand for these products can be significant, due to external factors like fashion trends, macroeconomic, and weather conditions.

Next, we extend the comparison to include expert Methods 4 and 5 as well, and therefore have to restrict the attention to the first assortment group in Season 3.

5.2. Advance demand information versus expert judgment

Table 5 gives the average MAPE, again for all SKUs with positive preview demand together (overall) and also per preview demand class. We remark that despite of the large differences in performance, the relatively small number of SKUs implies that almost none of these differences are significant (at the 5% significance level). For this reason, the significantly better methods (which would include most) are not indicated in Table 5, as they were for Table 2.

The two expert judgment methods provide similar performances, and overall clearly outperform all methods based on preview demand. Surprisingly, their comparative performance is especially good for SKUs with a large preview demand. This is counter-intuitive, as one would expect the statistical accuracy of the methods based on preview demand to increase with the preview demand size. More formally, assume that the scaling-up ratio of total demand (for the entire season) to preview demand has been correctly estimated based on historic sales of comparable SKUs. The inverse of this ratio, r, can be interpreted as the probability that a realized demand will be pre-ordered in the preview period. So, given total demand S, preview demand P follows a Binomial distribution with S repetitions and probability of success r. The associated mean and standard deviation of P are rS and $r(1-r)S$, respectively. Therefore, the scaled-up preview demand P/r has mean S (and is unbiased), variance $(1-r)S/r$ and squared coefficient of variation $(1-r)/rS$. Since the squared coefficient of variation is decreasing in S, SKUs with larger season demands and larger corresponding preview demands are expected to imply more reliable statistical extrapolations.

However, this is apparently not the case. Possible explanations are that preferences change during the season and that preview buyers, who order at a discount, are mainly price-sensitive customers and therefore only represent a segment of the market. Also, the pre-order catalogue is only distributed amongst loyal customers, who may not be representative of the entire customer base in the first place.

Table 5
Forecast errors (MAPE) averaged over the SKUs of Assortment group 1 for Season 3. All forecast methods are applied at the product group level.

		Advance demand information			Expert judgment	
		Method 1 (preview division) (%)	Method 2 (equal division) (%)	Method 3 (top-flop division) (%)	Method 4 (average) (%)	Method 5 (triangular distribution) (%)
$P_i = 0$	(13 SKUs)	100	56	57	88	96
$0 < P_i \leqslant 2$	(31 SKUs)	54	160	70	84	86
$2 < P_i \leqslant 5$	(25 SKUs)	58	142	101	41	41
$5 < P_i \leqslant 10$	(12 SKUs)	87	82	170	36	34
$P > 10$	(8 SKUs)	327	82	164	50	49
Overall $P > 0$	(76 SKUs)	89	133	106	59	59

6. Conclusion and recommendations

For a large set of SKUs and in two successive selling seasons, we have compared the accuracy of three quantitative forecasting methods based on advance (preview) demand information. The methods all use a top-down approach, and first forecast the aggregate total demand for a group of SKUs by scaling up the aggregate preview demands. They differ in the subsequent division of that aggregate forecast over the individual SKUs; proportional to preview demand (Method 1), equal (Method 2), or top-flop (Method 3).

A first result was that all top-down methods performed best when the top level was defined at the more detailed product group level rather than the broader assortment group level or even for all assortment groups together. This is a useful finding for the case company, as they currently use the assortment group level.

Amongst the considered top-down methods, the so-called top-flop method turned out to have the most robust performance. This method is based on the assumptions that, within a group of related SKUs, the top (flop) SKUs always account for a certain fraction of total demand for that group. To the best of our knowledge, this method has not been tested or described in the literature before. Our results certainly suggest that it would be interesting to test the method for other real-life cases as well.

An especially attractive feature of the top-flop method, as appeared from the overall results and was illustrated for specific product groups, is that it avoids over-forecasting for those SKUs with the largest preview demands. This is done by 'spreading' the total preview demand per class equally amongst the SKUs in a class, rather than dividing based on the individual preview demands per SKU as the more 'straightforward' Method 1 does. Over-forecasting leads to too large orders and obsolescence problems, where left over stock needs to be discarded or sold below the cost price at the end of the season.

For one assortment group with 89 SKUs, we extended the comparative study to include expert judgment methods. It turned out that these methods outperformed all other methods. However, Tukey tests showed the performance differences not to be significant at the 5% level. So, we should be cautious in announcing these expert judgment methods the 'winners'.

Our study has yielded several important insights for the case company. The company's existing approach (a combination of Methods 1 and 2) can be immediately improved on several accounts. Method 2 is outperformed by both Method 1 (preview division) and Method 3 (top-flop), and hence it is wise to exclude it from the existing approach. Further, our tests have shown that forecasting at the product group level produces more reliable estimates than forecasting at the higher assortment group level, as the company currently does. Additionally, we have provided an intuitive way of estimating group demand *M*, which does not carry the risks inherent in estimation by a committee of purchasers facing budget restrictions as is current practice (see Section 4.2). Although average forecast errors remain high, which is consistent with earlier findings [see, e.g.,](Chambers and Eglese, 1986; Gutgeld and Beyer, 1995; Fisher and Raman, 1996; Fisher and Raman, 1999; Fisher, 2002), small reductions in forecast error can potentially yield large increases in profits.

Our findings also show that alternative forecasting approaches are worthwhile to investigate. The top-flop method produces more robust estimates, in that it avoids the large forecast errors for SKUs with high preview demand that are produced by Method 1. In fact, the company currently investigates the use of Method 3 (top-flop) and considers hiring external consultants to implement forecasting using the top-flop logic (on a trial basis). The company is also considering to stop distributing pre-season catalogues, which would imply that preview demand data no longer become available. Our results suggest that forecasting based on expert judgment may indeed provide a good and perhaps even better alternative to methods based on preview demand, although we recommend further testing. If the company can abandon the preview exercise without jeopardizing the quality of the demand forecasts, it would realize a significant cost reduction and efficiency improvement. We remark that the top-flop logic can also be used in combination with expert judgments.

As discussed in Section 2, important shortcomings in the existing literature are that the proposed forecasting models are seldom tested on real data to show practical relevance, that they rely on demand data gathered using the selling season (a posteriori forecasting), and that they do not consider expert judgment. This research does not have any of these shortcomings. Our comparison includes both quantitative and expert judgment methods that rely on a priory information only, and is based on real data from an apparel company.

The forecasting methods discussed in this paper are generic and can be applied by any company that matches the following characteristics: (1) demand uncertainty is high and products are sold during a single selling season, and (2) the company cannot postpone ordering until actual demand information becomes available and can afford neither having significant leftovers at the end of a season, nor stockouts during the season. Demand data from previous seasons as well as some form of advance demand information and/or expert judgment data must be available. The comparative performance of the methods needs to be assessed on a case by case basis, and can be different for companies other than our case company.

There are two main limitations of this research, which also point to important directions for further research. First, only a subset of 89 SKUs was included for testing the expert judgment methods. This was done to limit the time involved for the group of experts, and thereby ensure that they stayed concentrated throughout the experiment. Future research could address further testing of expert judgment methods, ideally based on larger numbers of SKUs. Second, we did not examine market and economic conditions in depth. We did check that the ratio of total sales to preview demand was roughly the same for the three seasons in our data set, and found that this was indeed the case. However,

market and economic conditions could still play a role in affecting the preference for certain SKUs. Obviously, the correct identification of such conditions could be used to improve the methods that are based on advanced demand information. Further research could address this identification issue and test whether it leads to a significant improvement, also in comparison to methods based on expert judgment.

Finally, we remark that future research should not be restricted to the apparel industry. Similar forecasting problems to the one analyzed here can be found in the fashion, PC, mobile phones, consumer electronics, sporting goods, publishing and music industries. The methods and results presented here can therefore also yield valuable insights for forecasters in these industries, but more testing is certainly needed to confirm that.

References

Abernathy, F.H., Dunlop, J.T., Hammond, J.H., Weil, D., 2000. Control your inventory in a world of lean retailing. Harvard Business Review 78 (6), 169–176.

Au, K.-F., Choi, T.-M., Yu, Y., 2008. Fashion retail forecasting by evolutionary neural networks. International Journal of Production Economics 114, 615–630.

Bell, P.C., 2000. Forecasting demand variation when there are stockouts. Journal of the Operational Research Society 51, 358–363.

Blattberg, R.C., Hoch, S.J., 1990. Database models and managerial intuition: 50% model + 50% manager. Management Science 36 (8), 887–899.

Chambers, M.L., Eglese, R.W., 1986. Use of preview exercises to forecast demand for new lines in mail order. Journal of the Operational Research Society 37, 267–273.

Chambers, M.L., Eglese, R.W., 1988. Forecasting demand for mail order catalogue lines during the season. European Journal of Operational Research 34 (2), 131–138.

Chang, S.H., Fyffe, D.E., 1971. Estimation of forecast errors for seasonal-style-goods sales. Management Science 18 (2), B89–B96. Application Series.

Choi, T.-M., 2007. Pre-season stocking and pricing decisions for fashion retailers with multiple information updating. International Journal of Production Economics 106, 146–170.

M.L. Fisher, Rocket science retailing and the research opportunities it creates for us, September 23–27, 2002. Presentation at workshop 'The Role of Optimization in Supply Chain Management', Institute for Mathematics and its Applications (IMA). <http://www.ima.umn.edu/talks/workshops/9-23-27.2002/fisher/fisher.pdf>.

Fisher, M.L., Rajaram, K., 2000. Accurate retail testing of fashion merchandise: Methodology and application. Marketing Science 19 (3), 266–278.

Fisher, M.L., Raman, A., 1996. Reducing the cost of demand uncertainty through accurate response to early sales. Operations Research 44 (1), 87–99.

Fisher, M.L., Raman, A., 1999. Managing short-lifecycle products. <http://www.ascet.com/>.

Fisher, M.L., Hammond, J.H., Obermeyer, W.R., Raman, A., 1994. Coping with demand uncertainty at Sport Obermeyer. Harvard Business Review 72 (3), 90.

Fisher, M.L., Raman, A., Sheen McClelland, A., 2000. Rocket science retailing is almost here – Are you ready? Harvard Business Review (July–August), 115–124.

Fisher, M.L., Rajaram, K., Raman, A., 2001. Optimizing inventory replenishment of retail fashion products. Manufacturing and Service Operations Management 3 (3), 230–241.

Gallego, G., Ozer, O., 2001. Integrating replenishment decisions with advance demand information. Management Science 47 (10), 1344–1360.

Green, M., Harrison, P.J., 1973. Fashion forecasting for a mail order company using a Bayesian approach. Operational Research Quarterly 24 (2), 193–205.

Gutgeld, Y., Beyer, D., 1995. Are you going out of fashion? The McKinsey Quarterly 3, 55–65.

Hertz, D.B., Schaffir, K.H., 1960. A forecasting method for management of seasonal style-goods inventories. Operations Research 8 (1), 45–52.

Hill, R.M., 1997. Applying Bayesian methodology with a uniform prior to the single period inventory model. European Journal of Operational Research 98, 555–562.

Kogan, K., Herbon, A., 2008. Production under periodic demand update prior to a single selling season: A decomposition approach. European Journal of Operational Research 184, 133–146.

Kogan, K., Spiegel, U., 2008. Optimal policies for inventory usage, production and pricing of fashion goods over a selling season. Journal of the Operational Research Society 57, 304–315.

Lee, H.L., 2002. Aligning supply chain strategies with product uncertainties. California Management Review 44 (3), 105–119.

Mantrala, M.K., Rao, S., 2001. A decision-support system that helps retailers decide order quantities and markdowns for fashion goods. Interfaces 31 (3), S146–S165. Part 2 of 2.

Mostard, J., De Koster, R., Teunter, R., 2005. The distribution-free newsboy problem with resalable returns. International Journal of Production Economics 97 (3), 329–342.

Mostard, J.A.M., Teunter, R.H., 2006. The newsboy problem with resalable returns: A single period model and case study. European Journal of Operational Research 169 (1), 81–96.

Murray, G.R., Silver, E.A., 1966. A Bayesian analysis of the style goods inventory problem. Management Science 12 (11), 785–797.

Rahman, M., 2008. Stochastic demand forecast and inventory management of a seasonal product in a supply chain system. Ph.D. Thesis, Louisiana State University.

Raman, A., 1999. Managing inventory for fashion products. In: Tayur, S., Ganeshan, R., Magazine, M. (Eds.), Quantitative Models for Supply Chain Management. Kluwer Academic Publishers, pp. 791–805 (Chapter 25).

Sanders, N.R., Manrodt, K.B., 2003. The efficacy of using judgmental versus quantitative forecasting methods in practice. Omega, the International Journal of Management Science 31, 511–522.

Sethi, S., Yan, H., Zhang, H., 2003. Inventory models with fixed costs, forecast updates, and two delivery modes. Operations Research 51 (2), 321–328.

Sethi, S., Yan, H., Zhang, H., Zhou, J., 2007. A supply chain with a service requirement for each market signal. Production and Operations Management 16 (3), 322–342.

Tang, C.S., Rajaram, K., Alptekinoglu, A., Ou, J., 2004. The benefits of advance booking discount programs: Model and analysis. Management Science 50 (4), 465–478.

Thomassey, S., Happiette, M., 2007. A neural clustering and classification system for sales forecasting of new apparel items. Applied Soft Computing 7, 1177–1187.

Wang, T., Toktay, B.L., 2008. Inventory management with advance demand information and flexible delivery. Management Science 54 (4), 716–732.

Yan, H., Liu, KE., Hsu, A., 2003. Optimal ordering in a dual-supplier system with demand forecast updates. Production and Operations Management 12 (1), 30–45.