Accurate Retail Testing of Fashion Merchandise: Methodology and Application

Marshall Fisher • Kumar Rajaram
Operations and Information Management Department, The Wharton School, The University of Pennsylvania
Philadelphia, Pennsylvania 19104-6366, fisher@wharton.upenn.edu
Operations and Technology Management Area, The John E. Anderson Graduate School of Management
The University of California at Los Angeles, Los Angeles, California 90095-1481, kumar.rajaram@anderson.ucla.edu

Abstract

In a merchandise depth test, a retail chain introduces new products at a small sample of selected stores for a short period prior to the primary selling season and uses the observed sales to forecast demand for the entire chain. We describe a method for resolving two key questions in merchandise testing: (1) which stores to use for the test and (2) how to extrapolate from test sales to create a forecast of total season demand for each product for the chain. Our method uses sales history of products sold in a prior season, similar to those to be tested, to devise a testing program that would have been optimal if it had been applied to this historical sample. *Optimality* is defined as minimizing the cost of conducting the test, plus the cost of over- and understocking of the products whose supply is to be guided by the test.

To determine the best set of test stores, we apply a k-median model to cluster the stores of the chain based on a store similarity measure defined by sales history, and then choose one test store from each cluster. A linear programming model is used to fit a formula that is then used to predict total sales from test sales.

We applied our method at a large retailer that specializes in women's apparel and at two major shoe retailers, comparing results in each case to the existing process used by the apparel retailer and to some standard statistical approaches such as forward selection and backward elimination. We also tested a version of our method in which clustering was based on a combination of several store descriptors such as location, type of store, ethnicity of the neighborhood of location, total store sales, and average temperature of the store location. We found that relative to these other methods, our approach could significantly improve forecasts and reduce markdowns that result from excessive inventory, and lost margins resulting from stockouts. At the apparel retailer the improvement was enough to increase profits by more than 100%.

We believe that one reason our method outperforms the forward selection and backward elimination methods is that these methods seek to minimize squared errors, while our method optimizes the true cost of forecast errors. In addition, our approach, which is based purely on sales, outperforms descriptor variables because it is not always clear which are the best store descriptors and how best to combine them. However, the sales-based process is completely objective and directly corresponds to the retailer's objective of minimizing the understock and overstock costs of forecast error.

We examined the stores within each of the clusters formed by our method to identify common factors that might explain their similar sales patterns. The main factor was the similarity in climate within a cluster. This was followed by the ethnicity of the neighborhood where the store is located, and the type of store. We also found that, contrary to popular belief, store size and location had little impact on sales patterns.

In addition, this technique could also be used to determine the inventory allocation to individual stores within a cluster and to minimize lost demand resulting from inaccurate distribution across size. Finally, our method provides a logical framework for implementing micromerchandising, a practice followed by a significant number of retailers in which a unique assortment of merchandise is offered in each store (or a group of similar stores) tuned to maximize the appeal to customers of that store. Each cluster formed by our algorithm could be treated as a "virtual chain" within the larger chain, which is managed separately and in a consistent manner in terms of product mix, timing of delivery, advertising message, and store layout.

(*Merchandise Testing*; *Retailing*; *Mathematical Programming*)

Marketing Science © 2000 INFORMS
Vol. 19, No. 3, Summer 2000, pp. 266–278

0732-2399/00/1903/0266/$05.00
1526-548X electronic ISSN

1. Introduction

Retailers segment merchandise into basic and fashion products. Basic products have relatively stable demand and a long life-cycle, which make it fairly easy to forecast demand and manage inventory for a particular product using standard methods that rely on a sales history of the product. Forecasting and inventory management are much more difficult for fashion products. Their demand is highly unpredictable, and they have a short life-cycle—typically just a few months. They are often bought just once, at a time prior to the start of the actual sales season, and the decision of how much to buy is not based on actual sales of the product but merely on the subjective judgment of merchandisers and buyers about how well it will sell. We have found that these subjective forecasts have an average error of 50% or more. As a result, retailers frequently buy too little of some fashion products, resulting in lost sales and profit margin, and too much of other products, resulting in excess supply that must be marked down in price at the end of the season, frequently to the point where the product is sold at a loss.

To reduce these costly forecast errors, many retailers conduct experiments, called tests, in which products are offered for sale under carefully controlled conditions in a small number of stores. Tests are used to measure consumer reaction to a variety of variables including price, floor placement, marketing message, or some aspect of styling such as color, fabric, or silhouette. We focus here on a particular type of test, called a depth test, used to predict the season sales of a particular product. In a depth test, a supply of the product sufficient to avoid stockouts is placed in a small sample of stores for a two- to three-week period just prior to the start of the regular sales season. Sales in the test stores during this period are used to predict season sales for the chain, and this forecast is used as a basis for initial or replenishment orders. Although we focus specifically on depth testing in this paper, our method for choosing test stores could also be useful in other types of testing. For simplicity, we refer to this type of test as a *merchandise test*, or simply a *test*.

A retailer faces several issues in designing an effective merchandise test, including how many and in which specific stores to conduct the test and how to create a forecast for the entire chain based on test store sales. The decision of how many test stores to use must trade off the increased accuracy that comes from using more test stores against the cost of running the test, which is greater if more stores are used. The cost of running a test is incurred from administrative costs, the need to provide extra inventory to avoid stockouts during the test, possibly the cost of air-freighting merchandise to the test stores, and an opportunity cost on the store space used for the test, because test merchandise by its nature usually sells less well, on average, than regular merchandise. The high cost of testing generally leads retailers to use a small number of test stores (e.g., 5 to 25). Choosing a small sample of test stores from the hundreds of stores that comprise a large chain is challenging because of the variation in store characteristics such as location, climate, size, and demographics of the surrounding customer base.

Despite the practical relevance and complexity of this problem, we found nothing in the academic or managerial literature that describes how to design an effective merchandise test. There is extensive academic literature on test marketing (e.g., see Urban and Hauser 1980) that would appear to be relevant but turns out not to be directly applicable because it involves longer duration and observation of trials and assumes repeat purchases.

A number of articles that review current retail practice (Doyle and Gidengil 1977, Fox 1995, Hollander 1986, Pollack 1994, and Wilson et al. 1995) emphasize the importance of merchandise testing and highlight a need for more effective procedures, but they do not themselves describe how to conduct an effective merchandise test. Doyle and Gidengil (1977) review merchandise testing as part of the broader topic of retail experimentation and conclude that these methods, despite enormous potential, have thus far "made little contribution" to retailing because of both practical and theoretical problems.

We have been able to gather systematic information on testing practice as part of a broader, multiyear project involving 32 leading retailers of fashion type products, including apparel, computers, consumer electronics, entertainment software, books, music, toys, watches, and jewelry (see Fisher et al. 1999 for more details). Of the 27 retailers who answered the questions

on testing, 25 indicated that they conducted merchandise tests of some type, which supports that testing is widely used by retailers. Retailers were also asked to rate the effectiveness of their testing program on a 10-point scale, defining a 10 as the ability to predict sales from a test with an error of about 10%. The median answer was 6, suggesting that there is considerable room for improvement in testing accuracy in practice. In follow-up interviews, it appeared that the retailers who tested best had done an excellent job on many of the practical issues of testing but had not used any statistical methods to determine the most representative stores to use for a test or to interpret the results of a test. While retailers had diverse theories on how to test, to our knowledge none of them have subjected alternative approaches to a rigorous comparison to determine what actually works best.

This paper presents a methodology for resolving two key decisions in designing a merchandise test: in which stores to conduct the test, and how to create a season forecast for the chain based on test results. To choose test stores, we cluster the stores of the chain into groups that are similar based on historical sales of products that are similar to those to be tested. We apply linear programming to this sales history to estimate a formula to predict season sales for the chain from test store sales. We also consider clustering based on various store descriptor variables. Using data from three retailers—a specialty apparel retailer and two shoe retailers, Nine West and Meldisco—we compare the clustering methods to the existing process used by the apparel retailer and to two standard statistical approaches based on forecast accuracy and the cost of a supply plan. We find that sales-based clustering significantly outperforms the other methods on both metrics.

In the next section, we present the basic ideas behind our methodology. Section 3 presents the optimization models used to form clusters, select a test store within each cluster, and predict total season sales across all the stores of the chain from test store sales. Section 4 reports an application of these ideas to a retailer that specializes in women's fashion apparel, with over 1000 stores nationwide. Compared to the current testing process in place at the retailer, our method would reduce the cost of stockouts and of merchandise left over at the end of the season by enough to increase profit by 100%. In §5 we apply our method to data from two major shoe retailers, and obtain similar results. Section 6 presents conclusions to be drawn from our work and suggestions for future research.

2. Methodology

We describe a methodology for resolving the key decisions in designing a merchandise test: how many and specific stores in which to conduct the test, and how to create a season forecast for the entire chain based on test store sales. We assume that the retailer has (1) identified a set of products that they would like to test, (2) specified the time interval within which the products will be sold (the sales season), and (3) determined a test period during which the products will be offered for sale in selected stores to test their sales potential.

As mentioned in the introduction, retailers typically use a test period of two to three weeks just prior to the start of the regular season. The duration of the test presents the same trade-off as how many stores in which to test. A longer test is more accurate but more expensive. We have found two to three weeks to be typical practice, which seems to make sense. The test needs to be at least one week to control for interweek temporal effects. If the test occurs just prior to the start of the regular season, then the ending point of the test is constrained by the need to position supply based on the test by the start of the regular season. Hence, a longer test must start earlier, and the earlier the test begins, the greater the difference in conditions between the test period and the period when products will be sold.

We define the primary sales season as the period within the sales season during which the selling price of the product exceeds acquisition cost plus variable selling expenses, and the level of inventory at each store is sufficient to prevent supply shortages. We assume that the same price is charged at all stores during this period, although this common price may vary over time. It is important to restrict the sales data to the primary sales season; otherwise, sales below cost and supply shortages, both outcomes of bad planning, could distort the distribution of sales at individual stores. Note that this also ensures that sample sales are

not influenced by substitution because of a stockout of the product the customer was seeking. However, sample sales can be influenced by the dependency of demand with complementary or competitive products. If a store contains more complementary (competitive) products during the test than will be there during the regular selling season, then this will bias test sales upward (downward). This is an inevitable source of noise in a merchandise test, but it can be minimized by making the set of products offered during the test period as similar as possible to what will be offered during the regular season.

Because the purpose of merchandise testing is to create a forecast used to determine purchase quantities, it is important to understand the costs that result from forecast errors. If S_p is the actual demand for a product p during the primary season, U_p the per-unit cost of buying less than demanded, and O_p the per-unit cost of buying more than demanded, then the cost associated with forecast $\hat{S}_p$ is $U_p\max(S_p - \hat{S}_p, 0) + O_p\max(\hat{S}_p - S_p, 0)$. The understock cost U_p is often taken to be the profit contribution margin (price minus variable cost) that is lost if there is insufficient supply to meet demand. Excess supply is usually marked down in price at the end of the season and sold at a loss, so O_p is set to variable cost minus the marked down price.

Before describing our computational procedure, it is helpful to outline some features of retailing considered in designing our method. We can think of a product as being defined by a set of values for various attributes. Consumers differ in their preferences for attribute values, and hence the same product will have different appeal to different customers. In a retail chain with a large number of stores, consumer preferences will differ from store to store. While it is hard to directly measure the attribute preferences of customers that shop at a given store, given that attribute preference influences purchase decisions, the actual sales of a store can be thought of as a summary of the attribute preferences of customers at that store. In particular, if the percentage mix of products bought by customers at two different stores is similar, then we might infer that the customers of the two stores have similar preferences. This is the whole basis for micromerchandising, a practice followed by a significant number of retailers, in which a unique assortment of merchandise is offered in each store (or a group of similar stores) tuned to maximize the appeal to customers of that store (see Patterson 1995 and DiRomualdo 1998 for examples).

Our approach to testing is designed to recognize these features by using past sales of similar products to identify a set of test stores that collectively span the diverse segments of a large chain. In the next section, we first describe a procedure for choosing test stores, assuming that the number of test stores k has been specified. We then show how to set k to minimize the combined cost of testing and cost of forecast errors resulting from the test.

3. Model Description

We assume that the retailer has assembled a sales history of comparable products that were offered in a prior season or seasons. Appropriate comparable products are usually last year's products within the same classification. Let n denote the number of stores in the chain, m the number of previous products for which we have a sales history, S_{ip} the observed sales during the primary sales season at store i for product p, $S_p = \Sigma_{i=1}^{n} S_{ip}$, and $\bar{S}_{ip}$ the sales of product p in store i during a period comparable in timing and duration to a period during which a test would be conducted.

To choose k test stores, we partition the n stores of the chain into k clusters. The stores within each cluster are chosen to minimize a measure of dissimilarity based on the percentage of total sales represented by sales of each of the prior products in each store. Two stores that sold exactly the same percentage of each of the prior products would be in the same cluster, and all the stores within a cluster would sell approximately the same percentage of each of the prior products. We then choose a single test store within each cluster that best represents the cluster, in the sense that using test sales at this store to forecast sales of other stores in the cluster minimizes the cost of forecast errors.

We first describe the optimization model used to form clusters and select a test store within each cluster. This model is a specialized integer program known as the k-median problem, which we solve with the highly efficient algorithm given in Cornuejols et al. (1977).

Variables

$$y_j = \begin{cases} 1, & \text{if store } j \text{ is chosen as a test store,} \\ 0, & \text{otherwise;} \end{cases}$$

$$x_{ij} = \begin{cases} 1, \text{ if store } i \text{ is assigned to a cluster} \\ \quad \text{represented by test store } j, \\ 0, \text{ otherwise.} \end{cases}$$

Parameters

$I = (1, \ldots n) =$ store index set;

$P = (1, \ldots m) =$ prior product index set;

$$w_i = \sum_{p \in P} S_{ip};$$

$$\beta_{ip} = \frac{S_{ip}}{\sum_{p \in P} S_{ip}}, \; i \in I, \; p \in P;$$

$$d_{ij} = \sum_{p \in P} (U_p \max(\beta_{ip} - \beta_{jp}, 0) + O_p \max(\beta_{jp} - \beta_{ip}, 0)).$$

The Test Store Selection Problem (TSSP) is represented by the following integer program:

$$\text{Min} \sum_{i \in I} \sum_{j \in I} w_i d_{ij} x_{ij}, \tag{1}$$

subject to:

$$\sum_{i \in I} x_{ij} = 1, \; j \in I, \tag{2}$$

$$\sum_{j \in I} y_j = k, \tag{3}$$

$$0 \le x_{ij} \le y_j \le 1, \; i, j \in I, \tag{4}$$

$$x_{ij} \text{ and } y_j \text{ integral}, \; i, j \in I. \tag{5}$$

Equation (2) enforces the condition that each store is assigned to a test store, (3) that we have exactly k test stores, and (4) that stores are assigned only to chosen test stores. Objective Function (1) is structured to select k test stores that minimize the total overstock and understock costs if test sales at each test store are extrapolated to develop forecasts for stores assigned to that test store. To illustrate this point, let store i be represented by test store j (i.e., $x_{ij} = 1$). Then it would not be unreasonable to forecast sales for individual products in store i as $(w_i/w_j)S_{jp}$. The total cost associated with this forecast at this store is

$$\sum_{p=1}^{m} O_p \max\left(\frac{w_i}{w_j} S_{jp} - S_{ip}, 0\right) + U_p \max\left(S_{ip} - \frac{w_i}{w_j} S_{jp}, 0\right)$$

$$= \sum_{p=1}^{m} w_i \left(O_p \max\left(\frac{S_{jp}}{w_j} - \frac{S_{ip}}{w_i}, 0\right) + U_p \max\left(\frac{S_{ip}}{w_i} - \frac{S_{jp}}{w_j}, 0\right)\right)$$

$$= w_i \sum_{p=1}^{m} (O_p \max(\beta_{jp} - \beta_{ip}, 0) + U_p \max(\beta_{ip} - \beta_{jp}, 0)$$

$$= w_i d_{ij}.$$

Hence, the k-median problem can be interpreted as forming clusters and choosing a test store in each cluster that minimizes the cost of forecast errors, if total season sales at the test store were the predictor variable for its cluster. This also minimizes the expected cost of forecast errors based on test period sales, provided that the percentage of season sales that occur during the test period is approximately the same for all stores within a cluster. In this regard, it is worth noting that stores can differ not only in their sales mix but in the timing of their sales. If the timing of sales differs, then two stores might have an identical sales mix for the season but a different mix during the test period. In this instance, one of the stores might not be a good predictor for the other. As a practical matter, we have found that the most common cause of timing differences is climate. Southern stores sell spring/summer merchandise earlier than northern stores, and sell fall/winter merchandise later. We found in our testing that these climate differences also cause a difference in sales mix. Hence, while a timing difference that occurs separately from a mix difference is a potential problem, it did not occur in our test data. Were this to be an issue, we would recommend redefining d_{ij} to be based on the difference in the mix of season sales for store i and test period sales for store j, or to use a combined optimization approach defined in the Appendix.

We also tested a version of our method in which the d_{ij} were set based on a weighted combination of several store descriptors rather than on differences in stores' sales mix. The store descriptors were store latitude and longitude (the distance between two stores on this measure was simply the Euclidean distance between the stores), average temperature during the sales season, total store sales, ethnicity (percentage of white in the postal code where the store is located—not politically correct, but believed by many retailers to be indicative of sales patterns), neighborhood type (either urban, suburban or rural—we assigned a distance of 0 if the neighborhoods were the same and 1 if they were different), and store type (either mall, strip mall, or college campus—we assigned a distance of 0 if the store types were the same and 1 if different). The distance between two stores used in clustering was a weighted combination of the absolute difference of

these six measures, with nonnegative weights summing to 1 chosen to equalize the average influence of each of the six measures. We also tested a combined approach in which the distance between two stores was based on a combination of the six store descriptors and the difference in stores sales mix, with weights chosen to equalize the impact of these seven factors.

While we suggested above that w_i/w_j was a reasonable weight to use in predicting sales of a product at store i from its sales at store j, to ensure that we have the best possible weights and to adjust for the fact that test sales are for a shorter time interval than the full season, we use the following linear program (called the Test Sales Extrapolation Problem (TSEP)) to determine the optimal set of weights $(\alpha_1. \ \alpha_k)$, which scales test sales to estimate product sales for the entire season across the chain for a given set of k test stores.

$$Z_1(k) = \text{Min} \sum_{p \in P} U_p\theta_p + O_p\gamma_p,$$

$$\theta_p \geq \hat{S}_p - S_p \quad \forall p \in P,$$

$$\gamma_p \geq S_p - \hat{S}_p \quad \forall p \in P,$$

$$\hat{S}_p = \sum_{i=1}^{k} \alpha_i \bar{S}_{ip} \quad \forall p \in P,$$

$$\theta_p, \gamma_p \geq 0 \quad \forall p \in P.$$

In the discussion so far, we have assumed that the number of test stores k is given. To find the best number of test stores, we would choose k to minimize $C(k) = Z_1(k) + C_T k$, where C_T is the fixed cost for testing at each store. The fixed cost C_T results from the administrative cost of running the test at a store and the fact that test merchandise, by its nature, usually sells less well on average than regular merchandise; hence there is an opportunity loss on the shelf space dedicated to the test. Our procedure is fast enough that we can solve the optimization on k by enumeration of all values $k = 1, 2, \ldots$ noting that the largest number of test stores we need to consider can be bounded by $C(k)/C_T$ for any value of k for which we have evaluated $C(k)$, because using a number of stores larger than this would incur fixed costs greater than the total cost of a known solution.

It is possible to formulate a single optimization model to choose the identity of test stores and the weights in the linear forecast formula to minimize the total cost of overstock, understock, and testing. We tested this unified approach on the data described in §4 and found that it performed slightly worse than the approach described above. Further details of the unified approach are provided in the Appendix, including the formulation and our computational experience with it.

In the final analysis, the real test of our method is its ability to improve the accuracy of the merchandise testing process in an actual retail environment—a question we consider next.

4. Application at a Women's Specialty Apparel Retailer

We have tested these ideas with a large specialty women's apparel retailer with 1993 net sales that exceeded $1 billion, and with more than 1000 stores dispersed throughout the United States. In addition to varying by region, stores vary by size, format (mall, strip mall, etc.), urban vs. suburban, and the ethnicity of their target customer base. This retailer relied on testing as its primary tool for forecasting the demand for new products but also believed that the accuracy of its current methodology could be significantly improved, which is what led to our involvement.

In developing merchandise plans, the retailer divides the year into two seasons: fall/winter and spring/summer. The first week of the fall/winter season is the first full week of September, and the first week of the spring/summer season is the first full week in April. In their current methodology, they select 25 test stores whose total dollar sales are close to average store sales for the chain. Product tests are conducted at these stores over a three-week period. To develop a season forecast and supply plan for the entire chain, total sales during the test period are divided by two factors that are estimated from past sales history. The first factor equals the proportion of season sales that are historically observed during these three weeks; the second factor equals the proportion of season sales that are observed at these 25 stores.

To evaluate our approach and compare it to this approach, we considered the women's knit tops division, which historically represents one of the single largest

portions of investment and the highest level of sales uncertainty. The data available on which to evaluate our approach consisted of 1993 sales by store, by week, and by size of the 250 style/colors that make up the products of this division. To estimate the timing and duration of the primary sales season in this division, we analyzed 1993 sales data at the store/size level for these 250 style/colors. This analysis revealed that merchandise was in place at all stores by the start of week 3, that there was sufficient inventory at all the stores to meet potential demand through week 12, and that products were typically sold below cost after week 18. Consequently, we used a 10-week period, from the beginning of week 3 through the end of week 12, as the primary sales season.

We used these data to simulate the way our method would have performed if it had been used to design a test of this merchandise to be conducted during weeks 3 through 5. We used half of the 250 style/colors to fit the parameters of our model and the other half to test the accuracy of the model predictions. The planners at the knit top division classified the 250 style/colors into 16 product groups based on similarity in style and fabric texture. We selected 125 style/colors by choosing half the products in each of these groups. Let P denote this set of selected styles and $\bar{P}$ the remainder. For each $p \in P$, we used the sales data by store and product over the 10-week primary season to calculate w_i and d_{ij} and solved TSSP to optimality by the technique described in Cornuejols et al. (1977). For a given k, assume without loss of generality that the chosen test stores are indexed 1 through k. We then solved TSEP using the OSL solver in GAMS (see Brooke et al. 1992) to develop the linear function $\hat{S}_p = \Sigma_{i=1}^{k} \alpha_i \bar{S}_{ip}$, where $\bar{S}_{ip}$ is the actual sales of product p at store i during weeks 3 through 5, which is the period when a test would be conducted. For each $p \in \bar{P}$, we computed $\hat{S}_p = \Sigma_{i=1}^{k} \alpha_i \bar{S}_{ip}$, using $\bar{S}_{ip}$, the actual sales of product p at store i during weeks 3 through 5. We estimated the forecast error $\Sigma_{p\in\bar{P}} |\hat{S}_p - S_p|$, where S_p is the actual sales of product p across the chain during the primary season. Based on average selling price and costs during this period, we calculated the cost of these errors as the loss resulting from selling below cost and lost margin resulting from supply shortage, namely $\Sigma_{p\in\bar{P}} U_p \max(S_p - \hat{S}_p, 0) + O_p \max(\hat{S}_p - S_p, 0)$.

To provide a comparison, we also evaluated forecast errors and cost for the rules used by the planners at this retailer, the k median method based on store descriptors, alone and combined with sales mix differences, and two standard approaches to variable selection in linear regression, where the problem of choosing k test stores and a linear prediction function based on test sales at these stores is viewed as choosing the best k out of n possible variables in a linear regression. Given actual sales S_p and test sales $\bar{S}_{ip}$ for $i = 1, \ldots n$ and $p = 1, \ldots m$, we used the forward selection and backward elimination methods (Myers 1990) to choose k out of the n test store sales variables that best predict actual sales in the sense of minimizing the coefficient of determination R^2.

Notice that our method provides a sales forecast not only for the chain but for each cluster. The cluster forecasts were used to guide allocation of product to stores for the k-median clustering based on sales. In our prediction formula, $\alpha_p \bar{S}_{pj}$ represents the forecast of product j in the cluster of stores corresponding to test store p, and hence this quantity is the ideal amount to send to this cluster. Total sales volumes at individual stores were used as a basis to determine allocations to stores within a cluster through the formula $\alpha_p \bar{S}_{pj} w_i / \Sigma_{i\in I_p} w_i$, where I_p is the set of stores in the cluster represented by I_p. For other methods, we used the retailer's existing approach of allocating total supply in proportion to historical sales volume.

Forecast errors and costs for all six approaches for different values of k are shown in Table 1. Forecast errors and costs are an aggregate of these values determined at the store/style/color level. In aggregating forecast errors, we summed over all stores, styles, and colors the absolute difference between forecasted and actual sales and expressed it as a percentage of total actual sales. These results show that the k-median approach with clustering based on sales is significantly more accurate than the other methods. For example, for $k = 10$, it reduces costs resulting from forecast errors relative to the regression methods, the existing method in use at the retailer, or clustering based on store descriptors, by at least 8% of revenue. This improvement drops straight to the bottom line and would more than double profit when one considers that retailers typically earn profit before income tax of

Table 1 Forecast Error (F.E.) as a % of Unit Sales and Costs as % of Revenue for Different Methods

	k-Median Clustering Based on Sales		Regression with Forward Selection		Regression with Backward Elimination		Existing Method Used by Retailer		k-Median Clustering Based on Store Descriptors		k-Median Clustering Based on Store Descriptors and Sales	
k	F.E.	Costs	F.E.	Costs	F.E.	Costs	F.E.	Costs	F.E.	Costs	F.E.	Costs
1	36.0	40.0	37.9	42.2	38.6	43.0	71.0	74.0	39.8	44.3	38.2	42.4
5	17.0	25.9	27.0	35.9	28.0	37.0	52.1	58.0	23.3	32.8	20.3	30.6
10	12.9	20.9	19.1	29.2	20.0	30.1	41.9	46.0	19.0	28.9	16.7	23.0
15	12.0	19.5	17.1	26.0	16.8	23.7	38.1	42.3	16.8	23.2	15.9	22.4
16	11.7	19.3	16.8	23.6	16.6	23.0	37.7	41.9	16.5	22.9	15.4	22.2

about 3% to 5% of sales. The combined approach of clustering on store descriptors and sales mix difference comes close to, but is still dominated by, clustering on sales alone.

We believe that one reason our method outperforms the forward selection and backward elimination methods is these methods seek to minimize squared errors, while our method optimizes the true cost of forecast errors. In addition, our approach, which is based purely on sales, outperforms descriptor variables because it is not always clear which are the best store descriptors and how best to combine them. However, the sales-based process is completely objective and directly corresponds to the retailer's objective of minimizing the understock and overstock costs of forecast error.

To better understand the underlying causes of sales differences, it may be desirable to relate sales differences as much as possible back to store descriptors. To do this, we correlated store distance measures based on sales differences with the store descriptor variables temperature, ethnicity, location, store type, and size. We find that these descriptors as a group explain 85% to 89% of the variation in sales differences, with temperature by far the most significant variable. We also found that, contrary to popular belief, store size and location had little impact on sales patterns. Details of this analysis can be found in Fisher and Rajaram (2000) and in the online Appendix at ⟨http://mktsci.pubs.informs.org⟩.

We also conducted extensive analyses to address the following issues (details can also be found in Fisher and Rajaram 2000 and in the online Appendix): (1) assessing how the degree of collinearity between sales at the test stores affects the performance of these methods; (2) justifying our choice of 10 test stores to perform this analysis and evaluating the sensitivity of our method to the choice of the specific test store within a cluster; (3) the relationship between store size and test stores; and (4) the effect of temperature as an explanatory variable in the formation of the clusters.

Having evaluated our method on 1993 sales data, we also wanted to determine how well it would perform across multiple years because in actual use we would be fitting the model on sales that had occurred one year prior to actual introduction of the styles being tested. We obtained sales data for 30 style/colors from the 1994 fall season in the knit tops division that had been tested in 25 stores chosen by the planners. These were all new products that had not been on sale during 1993. We applied our clustering model as it was fit on the 1993 data to these (30) 1994 products. This exactly replicates how the model would be used in practice and hence is an accurate measure of its effectiveness.

Using the actual primary season sales (S_q) for these products in 1994, we computed total forecast error $\Sigma_{q\in Q} |\hat{S}_q - S_q|$ for all forecasts. Based on cost and selling price for these products during each week of 1994, we computed for all supply plans the loss resulting from selling below cost, and lost margin resulting from supply shortages. Forecast errors and

costs for all plans are shown in Table 2. The reduction in cost achieved by our plan relative to the other methods is 6.5% to 18%. This is better appreciated if we consider that the recorded net before tax profit in 1994 for these styles was 9% of revenue. Hence, our method could potentially improve profits by from $6.5/9 = 72\%$ to $18/9 = 200\%$.

The way we analyzed costs for these 30 style/colors treats demand as sales. However, because of stockouts, true demand usually exceeds sales. This generally biased comparisons against our method relative to the existing method because the inventory levels that conditioned sales had been determined by the existing method.

Assessing this effect is difficult because information on lost demand is not recorded. To estimate lost demand and margins, we first identified when every product/size combination in this group sold out in each store. We then defined the profitable season as the period during which products sold above cost, and we tabulated the percentage of profitable season sales that occurred each week for this group in total. These data were used to estimate sales that would have occurred after a stockout in a store/product/size combination before the end of the profitable season.

We applied our method for estimating lost demand to 104 style/color/size combinations in the knit series group and estimated that, in total, lost demand resulting from stockouts equaled 134% of sales actually realized. The lost margins resulting from this potential 134% increase in sales were around 57% of sales actually realized. Even if we allow for substitution between products to lower these estimates, the lost margins are clearly enormous.

Much of this lost demand was because of inaccurate distribution of inventory across sizes. Often we observed that a particular size in a given style/color accounted for a large proportion of stockouts, while other sizes of this style/color had to be eventually sold at markdowns below cost. To reduce such misallocation within the size distribution of a product, our clustering method could be applied to historical data on sales by size for a product to form clusters of products that had similar size selling patterns. We would then examine the nature of the products in each cluster as a way to understand how size distribution differs by product. For example, this might result in 10 distinct size distributions and a definition of the types of products that had a particular distribution. A new product could be assigned the size distribution of the product type that it best fits.

Recall that a test period of three weeks was used in developing the data reported in Table 2. To determine the impact of the test period length on cost and forecast error, we applied k-median clustering based on sales with test periods of varying lengths. Results are reported in Table 3 and suggest that the industry practice of a three-week test period is not unreasonable.

5. Application at Two Shoe Retailers

In this section we describe the application of our method at two major shoe retailers, Nine West and Meldisco. Nine West is the largest U.S. manufacturer of women's dress shoes and sells casual, career, and dress footwear and accessories worldwide through over 1500 of its own stores and 7000 locations in department, specialty, and independent shoe stores. We used our method to design a merchandise test for the spring sandals line, to be conducted during the first three weeks of the season. Using data on sales per week per store for 11 high fashion spring sandals sold at 140 stores for the first 26 weeks of 1997, we solved the TSSP for different numbers of test stores and used the sales during the first three weeks to find the alpha coefficients in the chain prediction formula.

With this formula and the chosen test stores, we then used sales for the first three weeks of the spring 1998 season to predict season sales for the 14 sandals in the 1998 spring line. Based on actual sales for these products in 1998, we computed forecast error and its associated understock and overstock costs, based on selling price, costs, and salvage value for each product. In computing costs, we assume that the initial order placed by the retailer is sufficient to cover the inventory requirements at stores from the start of the test to the time when the replenishment based on the revised forecast arrives.

We also applied all other methods described in our paper, except regression with variable elimination, which is computationally intensive and adds little ad-

Table 2 Forecast Error and Cost for Models Fit on 1993 Data and Applied to 30 1994 Products

	Existing Method with 25 Test Stores	k-Median Clustering Based on Sales with 10 Test Stores	Regression with Forward Selection with 10 Test Stores	Regression with Backward Elimination with 10 Test Stores	k-Median Clustering Based on Store Descriptors with 10 Test Stores	k-Median Clustering Based on Store Descriptors and Sales with 10 Test Stores
Forecast Error as % of Sales	38.0	19.5	30.0	31.0	29.0	25.5
Markdown Cost as % of Revenue	30.6	19.0	28.0	29.0	25.0	24.0
Lost Margin as % of Revenue	14.4	8.0	11.0	12.0	11.0	9.5
Total of Markdown Cost and Lost Margin	45.0	27.0	39.0	41.0	36.0	33.5

ditional information. Results are reported in Table 4. Note that clustering based on sales is superior to other methods, although clustering based on store descriptors combined with sales is a close second. Table 5 shows the impact of test period length on forecast error and cost for these data for *k*-median clustering based on sales.

Meldisco is a retailer that operates 1300 leased shoe departments in Kmart Stores. We applied our method to weekly sales data for 26 products for their 52-week season, assuming that the test takes place in the first six weeks and that a revised forecast is made based on this test. We did not have price and cost data for Meldisco, so we applied our method to minimize forecast error by setting the per unit costs of buying less than demanded equal to the cost per unit of buying more than demanded. In computing costs, we assume that the initial order placed by the retailer is sufficient to cover the inventory requirements at stores from the start of the test to the time when the replenishment based on the revised forecast arrives.

Table 3 Impact of Test Period Length on Forecast Error and Cost for k-Median Clustering Based on Sales

Length of Test Period	Forecast Error as % of Sales	Total Markdown Cost and Lost Margin as % Revenue
1	30.0	42.0
2	24.0	33.0
3	19.5	27.0
5	17.0	24.0
7	16.0	23.2
10	15.5	22.7

Table 4 Forecast Error (F.E.) as a % Unit Sales and Cost as % of Revenue for Different Methods applied to Nine West Data

	k-Median Clustering Based on Sales		Regression with Forward Selection		Heuristic Used by Another Retailer		k-Median Clustering Based on Store Descriptors		k-Median Clustering Based on Store Descriptors and Sales	
k	F.E.	Costs	F.E.	Costs	F.E.	Costs	F.E.	Costs	F.E.	Costs
1	49.1	45.3	56.2	51.8	61.0	56.2	53.7	49.4	51.1	46.6
5	18.9	18.1	27.5	26.8	29.8	28.5	23.5	22.9	20.7	20.0
10	17.6	16.6	26.0	24.9	27.9	27.1	22.1	21.1	19.1	18.2
15	17.0	15.4	24.7	24.1	27.3	26.8	21.2	20.3	18.5	17.6
20	16.7	15.0	24.6	23.9	26.8	25.9	20.8	19.8	18.2	17.3

Table 5 **Impact of Test Period Length on Forecast Error and Cost for k-Median Clustering Based on Sales for Nine West Data**

Length of Test Period	Forecast Error as % of Sales	Total Markdown Cost and Lost Margin as % Revenue
1	28.3	26.5
2	22.1	20.9
3	17.6	16.6
5	15.4	14.2
7	14.1	13.0
10	12.2	11.1
20	11.1	10.2
26	10.7	9.8

We used weekly sales for 13 of the products to select test stores and determine weights in the forecast formula. We then evaluated the test stores and forecast model on the remaining 13 products. We also applied all other methods, except regression, with variable elimination, which is computationally intensive and adds little additional information. Results are reported in Table 6. Note that our method is superior to others, although clustering based on store descriptors combined with sales is a close second. Table 7 shows the impact of test period length on forecast error for this data for *k*-median clustering based on sales.

6. Conclusions

Our goal in this paper is to expose the reader to an intellectually interesting problem context laden with opportunities for research that can have a high impact on retailer profits. The following are some conclusions to be drawn from the research reported here.

• The sales of a given product mix vary greatly among the stores of a large chain. Some, but not all, of this variation can be explained by store descriptors such as average temperature and ethnicity.

• A merchandise testing process that exploits this by clustering stores based on similarity of sales mix and choosing a single store within each cluster can provide forecasts of season demand for style/colors accurate to about 10% to 20%.

• This approach performs significantly better than alternative methods used in retail practice, based on standard statistical approaches or on clustering by store descriptor variables. The impact on cost of the superior performance is enough to double a retailer's profits.

The results reported here invite additional research on a number of topics.

• The unified approach described in the Appendix deserves further study.

• Choosing a small number of test stores from the many stores of a large chain is analogous to the statistical problem of choosing a parsimonious set of independent variables from a large potential set in a regression. The ability of the *k*-median approach to find variables that are minimally colinear, and hence more predictive, deserves study in this broader context.

• The relationship between the clusters formed by our algorithm and micromerchandising is worth exploring. Each cluster could be treated as a "virtual chain" within the larger chain, which is managed separately and in a consistent manner in terms of product

Table 6 **Forecast Error (F.E.) as a % of Unit Sales for Different Methods Applied to Meldisco Data**

k	Forecast Error for k-Median Clustering Based on Sales	Forecast Error for Regression with Forward Selection	Forecast Error for Heuristic Used by Another Retailer	Forecast Error for k-Median Clustering Based on Store Descriptors	Forecast Error for k-Median Clustering Based on Store Descriptors and Sales
1	41.8	50.7	54.1	45.9	43.7
5	13.7	21.2	24.3	17.2	14.9
10	12.7	20.4	22.7	16.4	13.9
15	12.2	19.2	21.6	16.1	13.6
20	12.0	18.3	21.3	15.6	13.4

Table 7 Impact of Test Period Length on Forecast Error for k-Median Clustering Based on Sales for Meldisco Data

Length of Test Period	Forecast Error as % of Sales
1	17.0
2	15.0
3	14.3
6	11.2
10	10.2
26	9.8
52	9.1

mix, timing of delivery, advertising message, store layout, etc. While this may require more managerial effort and careful attention to detail, the increasingly sophisticated information systems being installed by many retailers offer the opportunity to gather the required information and to automate much of the tedious data analysis tasks.

In conclusion, we believe that these trials at multiple retailers show that the method described here can greatly increase retailer profitability by improving the accuracy of merchandise testing.[1]

Appendix: Unified Optimization Approach

In this appendix we formulate an optimization model, which we call the Merchandise Testing Problem, to choose both the identity of k test stores and the weights of the linear forecast formula to minimize the total overstock and understock costs. As previously defined, n is the number of stores, $i \in I = (1, \ldots n)$ indexes the set of stores, m is the number of products for which we have sales history, $p \in P = (1, \ldots m)$ indexes the set of products, and

$\bar{S}_p$: sales for product p for the entire season across the chain,

$\bar{S}_{ip}$: sales of product p in store i during a period comparable in timing and duration to a period which a test would be conducted,

U_p: the understock cost per unit of buying less than demanded of product p,

O_p: the overstock cost per unit of buying more than demanded of product p, and

M: a sufficiently large real number. (In this application, it was set to 1000.)

Define the variables:

$$y_j = \begin{cases} 1, \text{ if store } j \text{ is chosen as a test store,} \\ 0, \text{ otherwise;} \end{cases}$$

$$x_{ij} = \begin{cases} 1, \text{ if store } i \text{ is assigned to a cluster represented by test store } j, \\ 0, \text{ otherwise;} \end{cases}$$

α_j: weight used to scale test sales at test store j to sales for the entire season across the chain;

θ_p: number of understock units of product p; and

γ_p: number of overstock units of product p.

The Merchandise Testing Problem (MTP) is given by the following mixed linear integer program:

$$Z_1(k) = \text{Min} \sum_{p=1}^{m} U_p\theta_p + O_p\gamma_p, \tag{1a}$$

$$\gamma_p \geq \hat{S}_p - \bar{S}_p \quad \forall p, \tag{2a}$$

$$\theta_p \geq \bar{S}_p - \hat{S}_p \quad \forall p, \tag{3a}$$

$$\theta_p, \gamma_p \geq 0 \quad \forall p, \tag{4a}$$

$$\hat{S}_p = \sum_{j=1}^{n} \alpha_j \bar{S}_{ip} \quad \forall p, \tag{5a}$$

$$-My_j \leq \alpha_j \leq My_j \quad \forall j, \tag{6a}$$

$$\sum_{j=1}^{n} y_j = k, \tag{7a}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \quad \forall i, \tag{8a}$$

$$0 \leq x_{ij} \leq y_j \leq 1 \quad \forall i, j, \tag{9a}$$

$$x_{ij}, y_j \in \{0, 1\} \quad \forall i, j. \tag{10a}$$

Objective function (1a) minimizes the total overstock and understock costs for the m products during the entire season across the chain; Equations (2a) and (3a) define overstock and understock quantities for each product, respectively; (4a) ensures that these values are nonnegative; (5a) estimates the forecast demand for each product by using test store sales at the k test stores; (6a) ensures that these weights are nonzero only if a store is chosen to be a test store; (7a) enforces the condition that we have exactly k test stores; (8a) that each store is assigned to a test store; and (9a) that stores are assigned only to chosen test stores. Integrality constraints are imposed by (10a).

Using data from the fashion apparel retailer, we solved MTP using the OSL solver in GAMS (see Brooke et al. 1992). Tables 1a and

[1]This research was supported in part by Alfred P. Sloan Foundation Grant 96-10-6 to the University of Pennsylvania. Our research has benefited from the advice and encouragement of Gerry Schleiffer (Vice President of Planning and Allocation of Nine West Group, Inc., at the time this study was conducted, and now Vice President of Planning and Allocation of United Retailer) and Jack Swem (Vice President of Planning and Distribution, Meldisco). We would like to thank Andreas Robotis for developing the software and performing the computations described in the paper and Professors Don Morrison and Dave Rubenstien for their useful comments in the earlier drafts of this paper. Finally, we are deeply appreciative to the two referees, the area editor, and Professor Brian T. Ratchford for their input during the review process, which has greatly increased the quality of this paper.

Table A1 **Forecast Error (F.E.) as a % of Sales and Cost as % of Revenue for the Combined Optimization Approach**

k	1	5	10	15	16
F.E.	37.1	18.6	14.3	13.2	13.0
Costs	41.2	26.3	22.1	21.2	20.9

Table A2 **Forecast Error and Cost for Combined Optimization Approach Fit on 1993 Data and Applied to 30 1994 Products**

Forecast Error as % of Sales	Markdown Cost as % of Revenue	Lost Margin as % of Revenue	Total of Markdown Cost and Lost Margin as % of Revenue
21.3	20	8.7	28.7

2a report forecast errors (F.E.) and costs for the cases in which results were reported in Tables 1 and 2. Comparing the results in Tables 1a and 2a with the results for other methods on the same cases, we see that the unified approach MTP did slightly worse than *k*-median clustering based on sales (1.2% higher cost on average for the cases in Table 1 and 1.7% higher cost for the case in Table 2), but better than all other methods.

It may seem surprising that a single optimization would not dominate the segmented approach computationally. The unified optimization MPT, by definition, can do no worse on the calibration sample than the segmented *k*-median clustering, but we found it performed worse on the evaluation samples. This could be because of a sampling error, although we believe the fact that *k*-median sales clustering uses full season sales in forming clusters results in a more robust choice of test stores. An advantage of the full optimization approach is that it better addresses a situation in which differences in timing of sales during the season are important to consider. Because of its greater simplicity and superior performance, we prefer the *k*-median sales clustering approach, but the differences between the two approaches are clearly small and the unified method deserves future research.

References

Brooke, A., D. Kendrick, A. Meeraus. 1992. *GAMS: A User's Guide.* The Scientific Press, San Francisco, CA.

Cornuejols, Gerard, Marshall L. Fisher, George L. Nemhauser. 1977. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Sci.* **23** 789–810.

DiRomualdo, Robert. 1998. The Borders system. Presented at First Annual Harvard/Wharton Merchandising Effectiveness Conference, May 20.

Doyle, Peter, B. Zeke Gidengil. 1977. A review of in-store experiments. *J. Retailing* **53**(2) 47–62.

Fisher, Marshall L., Kumar Rajaram. 2000. Accurate retail testing of fashion merchandise: Methodology and application. Working Paper, The Anderson School at UCLA.

——, Ananth Raman, Anna Sheen McClelland. 2000. Rocket science retailing is almost here—are you ready? *Harvard Bus. Rev.*, (July–August).

Fox, L. J. 1995. An integrated view of the assortment management process: The next frontier for leading retailers. *Chain Store Age Executive* (November).

Hollander Stanley C. 1986. A rearview mirror might help us drive forward. *J. Retailing* **62**(1) 7–10.

Myers, R. H. 1990. *Classical and Modern Regression with Applications.* PWS-KENT Publishing Company, Boston, MA.

Patterson, Gregory A. 1995. Target micromarkets its way to success: No 2 stores are alike. *Wall Street Journal* May 31.

Pollack, Elaine. 1994. Raising the bar: Keys to high performance. *Chain Store Executive Age.*

Urban, G. L., J. R. Hauser. 1980. *Design and Marketing of New Products.* Prentice-Hall Inc., Englewood Cliffs, NJ.

Wilson, B. L., M. Kingdom, T. Reeve. 1995. Quick hits in store level merchandise and inventory management. *Chain Store Age Executive* (November) 103–106.

This paper was received June 25, 1996, and was with the authors 36 months for 4 revisions; processed by Gary Lilien.