

Using Clustering to Improve Sales Forecasts in Retail Merchandizing

Mahesh Kumar

MSIS Dept., Rutgers Business School
Rutgers University, Newark, NJ 07102
maheshk@rutgers.edu

Abstract

Given sales forecasts for a set of items along with the standard deviation associated with each forecast, we propose a new method of combining forecasts using the concepts of clustering. Clusters of items are identified based on similarity in their sales forecasts and then a common forecast (or combined forecast) is computed for each cluster of items. The objective of clustering is to minimize the mean square error (MSE), which is the sum of the variance and squared bias of the combined forecasts. It is easy to show that combining forecasts from a group of items generally decreases the variance but increases the bias of the combined forecast. A new clustering method is proposed based on this tradeoff between the decreased variance and increased bias. A useful property of the proposed clustering method is that it automatically finds the right number of clusters. On a real dataset from a national retail chain we have found that the proposed method of combining forecasts produces significantly better sales forecasts than either the individual forecasts (forecasts without combining) or an alternate method of using a single combined forecast for all items in a product line sold by this retailer.

Keywords: Forecasting; Combining forecasts; Clustering; Retailing

1. Introduction

The ability to accurately forecast the sales volume for each item sold in each retail store is critical to the survival and growth of a retail chain because many operations decisions such as pricing, allocation and inventory management for an item are directly related to its sales forecast. However, most US retailers have large errors in their sales forecasts, and subsequently they lose large amounts of money due to mistakes on forecasting related decisions. According to analysis of retail and economic trends reported by the U.S. Census Bureau and National Retail Federation, the U.S. retailers are losing more than 200 billion dollars a year because of large forecast errors.

It is generally easier to obtain good forecasts for the aggregate sales from all items in a store or from all items in a product line than for each individual item in the store (See Mentzer and Cox, 1984 for a survey on how forecast errors

change as we go up in the product hierarchy). Mentzer and Cox suggest that the main reason why it is difficult to develop good sales forecasts for individual items is that often we have only few data points, sometimes only five or less data points, for each individual item.

In the event of insufficient data for each item, one solution is to pool data from a group of items and have a common forecast for all items in the group. The common forecast is expected to have smaller variance than the individual forecasts, albeit it may have higher bias. We know that the mean square error (MSE) of a forecast is equal to the sum of the variance and squared bias of the forecast. Therefore, if the bias created due to combining is not significant compared to the reduction in the variance of forecast, then the combined forecast would have lower MSE than the individual forecasts. Further, given a number of items, it is possible that combining only a subset (and not all) of items leads to lower MSE than individual forecasts. In that case, it is important to identify groups (or clusters) of items combining which leads to the lowest MSE. Once clusters of items are identified, one common forecast is computed for each cluster of items. In this paper, we propose a new clustering method that identifies clusters of items that minimize the total MSE of the combined forecasts. On a real data set from a large national retail chain, we have found that the proposed method of combining forecasts leads to lower MSE than the individual forecasts or an alternative method of computing one common forecast for all items in the data set, that is, having just one cluster consisting of all items.

The clustering method proposed here is an extension of the method proposed by Fisher (1958). Fisher's method, which also attempts to minimize the MSE, considers only the decrease in variance due to combining and does not account for the increase in bias. Since variance always decreases by combining, Fisher's method suggests that combining is always beneficial. In contrast, we propose a new clustering method based on the tradeoff between decreased variance and increased bias due to combining. The new method is able to identify when it is beneficial to combine and when it is not. A nice property of the proposed clustering method is that it automatically finds the right number of clusters in the data, which in itself is a challenging problem.

It is common to consider sales data in retailing as time series data. Duncan et al. (2001), Maharaj and Inder (1999), and Mitchell (2003) have proposed methods of clustering

time series data. All these methods have used similarity on historical data as the basis for clustering time series. We instead propose the use of similarity in the next period forecast and its variance to determine clusters of time series. The advantages of using the new approach are: (1) in many applications, we have access to only the next period forecast and its variance and not the entire history for each time series, (2) the next period forecast and its variance provide sufficient statistics if it is assumed that the forecasts are Gaussian distributed, and (3) the new method can handle judgemental forecasts which may not be based on the past data.

A large area of related research is the work on combining multiple forecasts for a single item. Starting with the seminal work by Bates and Granger (1969), a lot of work has been done in this area. See Clemen (1989) for a survey of the work on combining forecasts. Also see Menezes et al. (2000) for a more recent survey. These papers have shown that combining multiple unbiased forecasts for a single item leads to lower MSE than the individual forecasts. The basis for these papers is that each additional forecast decreases the variance of the combined forecast (by bringing additional information) without increasing its bias (because the combined forecast is also unbiased). This is not true when we combine unbiased forecasts for multiple items, because the combined forecast is no more unbiased. Therefore, it is necessary to develop new forecast combination methods that balance the tradeoff between decreased variance and increased bias when combining forecasts from multiple items.

The rest of the paper is organized as follows. Section 2 presents a new clustering model for combining forecasts. Section 3 presents two heuristic algorithms to solve the new clustering model. Section 4 presents an improved clustering model that accounts for the errors associated with each forecast. Section 5 provides a discussion on how the number of clusters changes as the amount of data available for analysis changes. Section 6 presents the empirical study results on a real-world data set. Finally, Section 7 provides a summary and directions for future research.

2. Simple Clustering Model

Let us consider n items indexed by $i = 1, \dots, n$. Let y_i be an unbiased forecast for item i . Let $E[y_i] = \mu_i$ and $Var(y_i) = \sigma_i^2$. Let s_i^2 be an estimate of σ_i^2 . Since individual forecasts are unbiased, the MSE for each item is equal to the variance of its forecast. The total MSE for n items is given by

$$MSE_{ind} = \sigma_1^2 + \dots + \sigma_n^2 \quad (1)$$

Suppose we combine these n items in a single group and have a common forecast for these items given by the average of n individual forecasts, equal to $\frac{y_1 + \dots + y_n}{n}$. It is easy to see that this common forecast, when applied to an individual item, is no more an unbiased forecast. The bias for item i is given by

$$Bias_i = \mu_i - \bar{\mu} \quad (2)$$

where

$$\bar{\mu} = E\left[\frac{y_1 + \dots + y_n}{n}\right] = \frac{\mu_1 + \dots + \mu_n}{n}. \quad (3)$$

Assuming that the individual forecasts are independent, the variance of the common forecast when used for item i is given by

$$Var_i = Var\left(\frac{y_1 + \dots + y_n}{n}\right) = \frac{\sigma_1^2 + \dots + \sigma_n^2}{n^2} \quad (4)$$

The total MSE for n items using the combined forecast is given by

$$\begin{aligned} MSE_{comb} &= \sum_i (Bias_i^2 + Var_i) \\ &= \sum_i \left[(\mu_i - \bar{\mu})^2 + \frac{\sum_i \sigma_i^2}{n^2} \right] \\ &= \sum_i \left[(\mu_i - \bar{\mu})^2 + \frac{\sigma_i^2}{n} \right] \end{aligned} \quad (5)$$

From Equations 1 and 5, we see that combining forecasts reduces the variance contribution towards total MSE from $\sum_i \sigma_i^2$ to $\frac{1}{n} \sum_i \sigma_i^2$, but increases the bias contribution from zero to $\sum_i (\mu_i - \bar{\mu})^2$. That means, if $\sum_i (\mu_i - \bar{\mu})^2$ is smaller than $(1 - \frac{1}{n}) \sum_i \sigma_i^2$, then the combined forecast has smaller MSE than individual forecasts and vice versa. In other words, if the items being considered have small difference in the means of their forecast ($\sum_i (\mu_i - \bar{\mu})^2$ is small) or have large variances in their forecast ($(1 - \frac{1}{n}) \sum_i \sigma_i^2$ is large), then treating these items as a single item and having only one forecast for all items instead of one for each item would lead to smaller MSE than the individual forecasts, even though the combined forecast would be biased.

We extend the above analysis to the case where we have $k \leq n$ combined forecasts. Each combined forecast corresponds to a cluster (or group) of items. Let k clusters be denoted by C_1, \dots, C_k . The k clusters form a non-overlapping (or disjoint) partition of n items. The total MSE for the items in cluster C_j is given by

$$MSE(C_j) = \sum_{i \in C_j} \left[(\mu_i - \bar{\mu}_{C_j})^2 + \frac{\sigma_i^2}{n_j} \right] \quad (6)$$

where $n_j = |C_j|$ and $\bar{\mu}_{C_j} = \frac{1}{n_j} \sum_{i \in C_j} \mu_i$. Total MSE for all k clusters is given by

$$MSE(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{i \in C_j} \left[(\mu_i - \bar{\mu}_{C_j})^2 + \frac{\sigma_i^2}{n_j} \right] \quad (7)$$

Our goal is to identify a set of clusters C_1, \dots, C_k that minimizes the total MSE in Equation 7. Unfortunately, this technique cannot be used directly, because μ_i and σ_i are unknown, and must be estimated from the data. The best estimates we have for μ_i is y_i and for σ_i^2 is s_i^2 . Replacing these in Equation 7, the objective of the clustering problem becomes

$$\min_{C_1, \dots, C_k} \sum_{j=1}^k \sum_{i \in C_j} \left[(y_i - c_j)^2 + \frac{s_i^2}{n_j} \right] \quad (8)$$

where $c_j = \frac{1}{n_j} \sum_{i \in C_j} y_i$.

This is a clustering problem on one-dimensional data. The objective function in this problem has two components.

The first component $\sum_{j=1}^k \sum_{i \in C_j} (y_i - c_j)^2$, which is same as the objective function of the standard one-dimensional clustering problem (Fisher, 1958), decreases as we decrease the number of data points in each cluster, and therefore it favors a large number of clusters. The other component $\sum_{j=1}^k \sum_{i \in C_j} \frac{s_i^2}{n_j}$ decreases as we increase the number of data points in each cluster, and therefore it favors a small number of clusters. This tradeoff between the first (bias) component and the second (variance) component of the objective function helps us obtain the correct number of clusters at the optimal point.

We have not been able to assess the computational complexity of the optimization problem in Equation 8. It would be an interesting research in itself to show either the problem is NP-hard or provide a polynomial time algorithm for the problem. In the absence of any such result, we present two heuristic algorithms that attempt to minimize the objective function of Equation 8.

3. Heuristic Algorithms

3.1 hClust: Hierarchical Greedy Heuristic

The *hClust* algorithm is based on a greedy heuristic, similar to the one used in Ward's agglomerative hierarchical clustering algorithm (Ward, 1963). The algorithm starts with n singleton clusters, each corresponding to a data point. At each stage of the algorithm, it merges a pair of clusters that leads to the maximum decrease in the objective function. The merging process stops when it is not possible to decrease the objective function any further. We show that the greedy heuristic at each stage of the *hClust* algorithm is equivalent to combining the most similar pair of clusters according to a *similarity function* defined below.

Proposition 1 *At each step of the hClust algorithm, we merge a pair of clusters C_u and C_v for which similarity*

$$sim(u, v) = \frac{n_u n_v}{n_u + n_v} [\bar{s}_u^2 + \bar{s}_v^2 - (c_u - c_v)^2] \quad (9)$$

is maximized, where $\bar{s}_j^2 = \frac{\sum_{i \in C_j} s_i^2}{n_j}$, for $j = u, v$.

Proof : Let E_j be the contribution from cluster C_j to the objective function, i.e.,

$$E_j = \sum_{i \in C_j} [(y_i - c_j)^2 + \frac{s_i^2}{n_j}], \quad j = 1, \dots, k \quad (10)$$

Suppose we choose to merge clusters C_u and C_v during an iteration of *hClust*, and let the resulting cluster be C_w . The net decrease in the objective function is given by

$\Delta E_{uv} = E_u + E_v - E_w$. Replacing the values of E_u , E_v , and E_w using Equation (10), and simplifying the resulting expression gives

$$\Delta E_{uv} = \frac{n_u n_v}{n_u + n_v} [\bar{s}_u^2 + \bar{s}_v^2 - (c_u - c_v)^2] \quad (11)$$

Maximizing ΔE_{uv} is, therefore, same as maximizing similarity $sim(u, v) = \frac{n_u n_v}{n_u + n_v} [\bar{s}_u^2 + \bar{s}_v^2 - (c_u - c_v)^2]$ among all possible pairs of clusters, C_u and C_v . ■

It is useful to note that $sim(u, v)$ can take both negative as well as positive values. The hClust algorithm combines clusters that have highest similarity and continues combining as long as there are pairs of clusters with positive similarity. The time complexity of hClust algorithm is $O(n^2)$. An important note to make here is that while most hierarchical clustering methods require a threshold (which is often decided subjectively) as a stopping criterion for the hierarchical algorithm, the stopping threshold for hClust is always zero. The hClust algorithm is formally described in Algorithm 1.

Algorithm 1 : *hClust*($y_1, \dots, y_n; s_1, \dots, s_n$)

- 1: **initialization:**
 - 2: $C_i = \{x_i\}, i = 1, \dots, n;$
 - 3: Calculate similarity between all pairs of clusters;
 - 4: For each cluster, record its most similar cluster;
 - 5: **end initialization**
 - 6: $maxsim = 1;$
 - 7: **while** $maxsim > 0$ **do**
 - 8: Find a pair of clusters C_u and C_v with highest similarity;
 - 9: Merge C_u and C_v into one cluster $C_w = C_u \cup C_v;$
 - 10: Calculate similarity of C_w to all other clusters;
 - 11: For each cluster, update its most similar cluster (in case C_w becomes its most similar cluster);
 - 12: $maxsim =$ similarity between the most similar pair of clusters;
 - 13: **end while**
 - 14: **end while**
-

3.2 kClust: Contiguous Clustering Heuristic

The basis for kClust algorithm is the following result by Fisher (1958). In the special case, when all s_i are zero, then for a given value of k , the optimal solution for the problem in Equation 8 is achieved by a *contiguous clustering* defined below.

Definition 1 *A clustering is called contiguous if it satisfies the following condition: for any triplet of data points x_u, x_v and x_w that have the order $x_u < x_v < x_w$, if points x_u and x_w belong to the same cluster, then x_v must also belong to that same cluster.*

Fisher's result implies that, in the special case when all s_i are zero, we can find an optimal clustering using dynamic programming in polynomial time. The contiguous clustering property does not hold in general if s_i 's can take non-zero values. It is not even clear if there exists a polynomial

time algorithm to find an optimal solution for the proposed clustering problem. We propose a heuristic solution that considers all possible contiguous clustering for all choices of k and picks the one that minimizes the objective function. This can be done in polynomial time using dynamic programming. A pseudo-code for the proposed heuristic is described below, whose time complexity is $O(n^3)$.

Algorithm 2 : $kClust(y_1, \dots, y_n; s_1, \dots, s_n)$

- 1: Sort y 's in increasing order;
 - 2: Let $w(i, j)$ be the contribution of cluster $\{y_i, \dots, y_j\}$ to the objective function;
 - 3: $w(i, j) = \sum_{r=i}^j [(y_r - c_{ij})^2 + \frac{s_r^2}{j-i+1}]$, where $c_{ij} = \frac{\sum_{r=i}^j y_r}{j-i+1}$, $1 \leq i < j \leq n$;
 - 4: Let $f(m, k)$ be the best contiguous clustering of points $\{y_1, \dots, y_m\}$ into k clusters;
 - 5: $f(m, 1) = w(1, m)$ for all $m = 1, \dots, n$;
 - 6: $f(m, k) = \min_{k-1 \leq i < m} (f(i, k-1) + w(i+1, m))$ for all $1 \leq k \leq m \leq n$;
 - 7: Best contiguous clustering is the one that minimize $\min_{1 \leq k \leq n} (f(n, k))$;
-

In the empirical study presented in Section 5, we found that in all of the instances of this study, kClust produced clustering solutions with lower objective values than the ones produced by hClust. Thus, while kClust takes more computational time than hClust, it produces better solution than hClust.

4. Weighted Clustering

There are two shortcomings in the clustering model proposed above. First, the common forecast for a cluster of items is computed as the simple average forecasts for items in that cluster, that is, each item is given equal weight in the average forecast. Bates and Granger (1969) have shown that, in the presence of variance statistics available for each forecast, it is better to use a weighted average common forecast, where the weight on a forecast is the inverse of its variance. The weighted average forecast for items in cluster C_j is given by

$$c'_j = \frac{\sum_{i \in C_j} \frac{y_i}{s_i^2}}{\sum_{i \in C_j} \frac{1}{s_i^2}} \quad (12)$$

whose variance is equal to

$$Var(c'_j) = Var\left(\frac{\sum_{i \in C_j} \frac{y_i}{s_i^2}}{\sum_{i \in C_j} \frac{1}{s_i^2}}\right) = \frac{1}{\sum_{i \in C_j} \frac{1}{s_i^2}}. \quad (13)$$

The second shortcoming of simple clustering model is that its objective criterion is dominated by items that have large variance in their forecast. In order to balance each item's contribution to the objective function, we modify the objective function to minimizing the total weighted mean square error (WMSE) as defined below.

$$WMSE = \sum_{i=1}^n \frac{1}{s_i^2} MSE_i \quad (14)$$

where MSE_i is the MSE contribution from the i^{th} item to the objective function. With these two modifications the new clustering criterion becomes

$$\min_{C_1, \dots, C_k} \sum_{j=1}^k \sum_{i \in C_j} \frac{1}{s_i^2} [(y_i - c'_j)^2 + \frac{1}{\sum_{i \in C_j} \frac{1}{s_i^2}}] \quad (15)$$

An important advantage of weighted clustering is that its optimal solution is always a contiguous clustering, and therefore, it is solvable in polynomial time using dynamic programming.

Proposition 2 *Weighted clustering always has an optimal solution given by a contiguous clustering.*

Proof : The contribution from the variance term in weighted clustering is

$$\sum_{j=1}^k \sum_{i \in C_j} \frac{1}{s_i^2} \frac{1}{\sum_{i \in C_j} \frac{1}{s_i^2}} = \sum_{j=1}^k \frac{\sum_{i \in C_j} \frac{1}{s_i^2}}{\sum_{i \in C_j} \frac{1}{s_i^2}} = k, \quad (16)$$

which is a constant for a fixed k . Therefore, for a fixed k , Equation 15 is equivalent to

$$\min_{C_1, \dots, C_k} \sum_{j=1}^k \sum_{i \in C_j} \frac{1}{s_i^2} (y_i - c'_j)^2 \quad (17)$$

which always has a contiguous optimal solution (Fisher, 1958). ■

Corollary 1 *Weighted clustering is solvable in polynomial time.*

Proof : Using the contiguous optimal property of weighted clustering, one can find the best k clusters for each value of $1 \leq k \leq n$ using dynamic programming. Then an optimal weighted clustering corresponds to the value of k for which the best k clusters have the lowest objective value. This can be done in $O(n^3)$ time in the same way as in Algorithm 2 for kClust. ■

5. Empirical Study

In order to assess the usefulness of the proposed method in practice, we carried out an empirical study on real data from a major national retail chain. We obtained one year of weekly sales data from April 13, 2002 to April 12, 2003 for 764 items of women's summer clothing. Not all of these items were sold for the whole of the year. We found that a large number of items were introduced at the beginning of week 10 (April 15-21) and were taken off the shelf at the end of week 30 (November 2-8). In order to keep the empirical study simple, we considered only those items that were sold for all of this 21-week period from week 10 to week 30. This left us with 411 items. We found negative sales during several weeks. This could be due to return merchandise or data entry mistakes. There was no way to verify this. In practice, retailers either convert the negative entries into zero or leave them as they are. We chose the second option.

	Avg. MSE using Individual Forecasts	Avg. MSE using Simple Average	% decrease in MSE of simple average over individual forecasts	Avg. MSE using Weighted Average	% decrease in MSE of weighted average over individual forecasts
Moving Average	102.43	131.37	-28.3 %	129.13	-26.1 %
Exp. smoothing	125.27	169.55	-35.3 %	158.24	-26.3 %
Box-Jenkins	131.16	162.71	-24.1 %	159.42	-21.5 %

Table 1. Empirical results using simple and weighted averages

	Avg. MSE using Individual Forecasts	Avg. MSE using Simple Clustering	% decrease in MSE of simple clustering over individual forecasts	Avg. MSE using Weighted Clustering	% decrease in MSE of weighted clustering over individual forecasts
Moving Average	102.43	94.62	7.6 %	89.97	12.2 %
Exp. smoothing	125.27	105.93	15.4 %	101.36	19.1 %
Box-Jenkins	131.16	106.86	18.5 %	98.02	25.3 %

Table 2. Empirical results using simple and weighted clustering

We obtained the merchandize id (also called class id or product line id) for each item. The merchandize id refers to a group of merchandize that have similar characteristics, for example, all summer shorts (or all winter boots) would have the same merchandize id. It is generally believed that merchandize id provides a good classification of products sold by a retailer, and therefore forecasts from items having different merchandize id should not be combined. The retailer whose data we analyzed generally uses the individual forecast of an item for its operations planing. But if it is not possible to use the individual forecast due to insufficient data (or lack of confidence in the individual forecast) then they replace an item’s forecast by the common forecast of the merchandize id to which this item belongs. The common forecast of a merchandize id is computed as the simple average of individual forecasts of items in that merchandize id. In our dataset there were 11 unique merchandize id. The number of items in a merchandize id ranged from 10 to 83.

We chose this particular data set for the empirical study because there was no price change for any item during the entire selling season. This simplified our analysis. We found that the sales data showed significant seasonal behavior, which must be estimated in order to forecast future sales. We collected previous year’s sales data for all items and estimated a common seasonality for these items using the method proposed by Kumar et al. (2002). The estimated seasonality consists of 52 indices, one for each week. The weekly sales data for each item was adjusted for seasonality by dividing it by the corresponding seasonality indices. The individual sales forecast was then computed for each item using the adjusted sales data. We experimented with three different forecasting models: moving average, exponential smoothing, and Box-Jenkins method. For moving average, we used 4-weeks moving average, and for expo-

ponential smoothing, we used smoothing constant $\alpha = 1/3$, which are commonly used parameters in practice by a number of retailers.

Given that we are at time t , first we calculate the next period individual forecast for each item in a merchandize id. Then we compute combined forecasts for groups of items in a merchandise id using both clustering methods: simple clustering and weighted clustering. Note that we apply clustering to each merchandize id separately, that is, we never combine forecasts from items belonging to different merchandize id. For simple clustering, we run both hClust and kClust algorithms and pick the one that gives lower value for the objective function of simple clustering. We compute individual forecasts and combined forecasts for time points $t = T_1$ to T_2 . In this example, we take $T_1 = 5$ because forecasts for 4-weeks moving average are not available before week 5, and $T_2 = 21$ because we have a total of 21 weeks of data for each item. We report the *Average MSE* computed as below.

$$\text{Average MSE} = \frac{\sum_{i=1}^N \sum_{t=T_1}^{T_2} (\text{forecast}_{it} - \text{actual}_{it})^2}{N * (T_2 - T_1 + 1)} \quad (18)$$

where forecast_{it} is the sales forecast for item i at time t , actual_{it} is the actual sales for this item at time t , and N is the total number of items. The summary results are reported in Tables 1 for all three forecasting models.

We found that, for all three forecasting models, simple clustering had lower Average MSE than the individual forecasts, and weighted clustering had even lower Average MSE. The reduction in Average MSE by simple clustering was in the range of 7.6 % to 18.5 %, and the reduction by weighted clustering was in the range of 12.2 % to 25.3 %. We also found that, among three methods we experimented

with, 4-weeks moving average performed the best, which is also the method being used by this retailer in practice.

We also considered an alternate method of combining forecasts where all items belonging to a merchandise id are grouped into a single cluster, that is, we have one cluster for each merchandise id. The common forecast for each merchandise id is computed in two ways: simple average and weighted average of individual forecasts for all items in the merchandise. In weighted average, each item gets a weight equal to the inverse of its variance, as in Equation 12. The summary results are reported in Tables 2 for this experiment.

We found that this alternate method of combining always produced worse forecasts than individual forecasts. The Average MSE increased by more than 20 % for all three methods. Interestingly, weighted average performed slightly better than simple average (which is consistent with the results in Table 1), but still it was a lot worse than individual forecasts.

The results in Tables 1 and 2 suggest that, if the individual forecasts are not satisfactory (due to high MSE), the current practice of replacing an item's individual forecast by the common forecast of the merchandise id to which the item belongs to may not be the right thing to do. Results in Table 2 show that the current practice may produce worse forecast (on the measure of Average MSE) than the individual forecasts. It would be better to identify clusters of items within each merchandise id and have a common forecast for each cluster rather than having a common forecast for the merchandise id.

6. Summary and Future Research Directions

In this paper we have proposed a new method of combining forecasts from a set of items using the concepts of clustering. We formulated a new clustering problem, called simple clustering, that attempts to minimize the total MSE of forecasts. Simple clustering is based on the tradeoff between increased bias and decreased variance when forecasts from multiple items are combined. We are not able to show whether simple clustering is solvable in polynomial time or not. We have proposed two heuristic algorithms that attempt to minimize the objective function of simple clustering. We have also proposed a different clustering method, called weighted clustering, that gives each forecast a weight equal to the inverse of its variance. A nice property of weighted clustering is that it is solvable in polynomial time. In an empirical study with sales data from a national retail chain, we found that simple clustering reduced the overall MSE of forecasts by 7.6 % to 18.5 %, and weighted clustering reduced it further by another 3.7 % to 6.8 %.

We see this paper as the first work on combining forecasts from multiple items using clustering. The paper leaves a number of open questions for future research.

We have assumed that individual forecasts for all items are independent. It would be useful to study how the proposed method changes in the presence of correlation information available for pairs of individual forecasts.

The weighted clustering model uses the inverse of variances of forecasts as weights. There has been much work done on how to find the optimal weights for combining forecasts. One could extend weighted clustering by making weights as decision variables.

Empirically we found that weighted clustering always performed better than simple clustering. It is not clear if it was because of our inability to find an optimal solution for simple clustering or because weighted clustering is indeed a better clustering method. In either case, finding an optimal solution (or a better heuristic) for simple clustering is an interesting research in itself.

7. Acknowledgment

The author would like to thank Nitin R. Patel for research discussions, ProfitLogic Inc. for providing data, and Research Resource Committee at Rutgers Business School for financial support.

References

- [1] Bates JM, Granger CWJ. The combination of forecasts. *Operational Research Quart.* 1969; 451-468.
- [2] Clemen R. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 1989; 5; 559-583.
- [3] Duncan G, Gorr WL, Szczypula J. Forecasting analogous time series. In: Armstrong JS (Edn), *Principles of forecasting: A handbook for researchers and practitioners*, Kluwer Publishers; 2001.
- [4] Fisher WD. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 1958; 53, 789-798.
- [5] Kumar M, Patel NR, Woo J. Clustering seasonality patterns in the presence of errors. *Proceedings of KDD 2002*; 557-563.
- [6] Maharaj EA, Inder BA. Forecasting time series from clusters. *International Symposium on Forecasting* 1999.
- [7] Menezes LM, Bunn DW, Taylor JW. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research* 2000; 120; 190-204.
- [8] Mentzer JT, Cox JE. Familiarity, application, and performance of sales forecasting techniques. *Journal of Forecasting* 1984; 3(1); 27-36.
- [9] Mitchell RJ. Forecasting electricity demand using clustering. *Proc of the 21st IASTED International Conference on Applied Informatics* 2003; 225-230.
- [10] Ward JH. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 1963; 58; 236-244.