Demand Forecasting
of New Products
Using Attribute Analysis

Marina Kang

A thesis submitted in partial fulfillment
of the requirements for the degree of

BACHELOR OF APPLIED SCIENCE

Supervisors:
Professor W.L Cleghorn
Professor J.K Mills

Industry Partner:
Teradata Corporation

Industry Supervisor:
Ed Kim

Department of Mechanical and Industrial Engineering
University of Toronto

April 1, 2009

Abstract

This thesis is a study into the demand forecasting of new products (also referred to as Stock Keeping Unit or SKU) employing attribute analysis techniques. The objective of this thesis is to improve upon currently employed new-SKU demand forecasting methods which involve the processing of large amounts of historical demand data taken from seemingly similar existing SKUs. The objective is accomplished by way of attribute analysis techniques that serve to refine the data set of existing SKUs, from which the new SKU forecast is drawn, to encompass only data that truly reflect the new SKU demand.

The first approach utilizes Conjoint Analysis (CJA) to quantitatively identify existing SKUs that are similar to the new SKU based on consumer demand. Forecasts based on the CJA results produced forecasts with improved accuracies over current methods, with overall forecast accuracies of up to 12.96%.

An alternate approach to refining the original data set identifies and isolates trends in the historical data. The new SKU forecast is generated by assimilating one of the identified trends. Principal Component Analysis (PCA), a method used to identify relationships and patterns that exist in data of large dimensions, is utilized to identify demand trends across large quantities of SKUs. The PCA results were positive with demand trends clearly emerging from within the data set as a whole. However, although PCA based forecasts showed some improvement at 15.14% accuracy, this was not a statistically significant improvement over the CJA forecasts.

Acknowledgements

The author would like to thank Mr. Ed Kim of Teradata Corporation for his constant and unfailing support and time during this thesis project. Beyond providing the raw data upon which all analysis was conducted, none of this would have been possible without the vast amounts of knowledge, insight and direction he has graciously provided throughout the course of this project.

Acknowledgment and gratitude is also extended to the Ontario Centres of Excellence, whose support and funding has been of great assistance to the completion of this project. Additionally, the author would like to thank Professors W.L Cleghorn and J.K Mills for providing the great opportunity to connect with an industry partner to complete a thesis project through the Connections Program.

The author would like to express her great appreciation for the tremendous dedication shown by all the aforementioned people throughout the duration of this thesis, without whose support, this project would not have come into fruition.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Demand forecasting (hereon in referred to simply as Forecasting) aims to predict the amount of product or service to be purchased by consumers [1]. In retail settings, accurate time series forecasts result in, for example, fewer stock outs and savings in ordering and stock keeping (inventory) costs. A combination these factors translate into large savings for the company and customer satisfaction.

The driving force behind accurate forecasting lies in the rigorous analysis of the product's historical data. In this way, the field of demand forecasting is vast and requires the consideration of many factors such as cost, promotions, cannibalization, season and weather to name a few. It is essential to accumulate sufficient amounts of historical data and consider the effects of all the significant factors driving the demand in order to produce accurate and reliable forecasts.

An underdeveloped field within the scope of forecasting is the forecasting of new products (also referred to as Stock Keeping Units or SKUs) that enter the market. Although new SKUs do not have historical demand information to offer any insight into its initial forecasts, current forecasting models do employ the use of historical demand of very loosely defined similar existing products or SKUs. These models, however, are overly simplistic and do not necessarily employ the same degree of rigorous data analysis of the existing SKU historical demand to generate new SKU forecasts in comparison to the analysis performed when forecasting for SKUs already in market.

One current new SKU forecasting model allows the new SKUs to assimilate the demand of a similar existing SKU, chosen by the forecaster [2]. This method, in addition to being a manual process, leads to unreliable forecasts as there may be underlying obscure SKU attributes overlooked by the forecaster when choosing a “similar” SKU demand to assimilate.

Alternatively, another process involves looking into the class of SKUs to which the new product belongs and averaging the previous year’s demand for SKUs in that class as a basis for the demand of the new SKU. As will be shown, however, this method leads to widely inaccurate forecasts as demand greatly varies, even among SKUs within the same class.

1.1 Motivation

The motivation behind this thesis, therefore, is to help fill the void in new SKU demand forecasting by applying more rigorous data analysis techniques using the historical demand of existing SKUs in the generation of new SKU forecasts.

1.2 Objectives

The objective of this thesis is to apply statistical techniques to the forecasting of new SKUs in the hopes of producing more accurate forecasts over current methods. Current forecasting methods are based on averaging the historical demand of large data sets of similar existing SKUs, where the term “similar” is very loosely defined or subjectively by the business manager. Alternate methods explored in this thesis still involve averaging the historical demand of existing SKUs to obtain new SKU forecasts; however, fastidious attention is given to the data set over which the averaging occurs. Specifically, two statistical techniques, Conjoint Analysis and Principal Component Analysis, are used to selectively choose the existing SKUs that appear in the refined data sets used for new SKU forecasting.

1.3 Hypothesis

Current demand forecasting methods can be greatly enhanced with the statistical techniques that provide insight into SKU similarities and historical demand trends. It is hypothesized that Conjoint Analysis and Principal Component Analysis concepts applied to the large data sets of existing SKUs will generate a significantly smaller selective data set of only relevant SKUs from which more accurate forecasts can be produced.

Conjoint Analysis (CJA) is used to assign importance weights, derived from consumer demand, to SKU attributes [3]. The theory is that comparisons made between existing and new SKUs based on the importance weights will identify which existing SKUs the new SKU is most similar to. The results are then used to compile a refined data set of quantified similar existing SKUs, from which the new SKU forecast is generated.

Principal Component Analysis (PCA) is a statistical tool that is used to identify and isolate patterns in data [4]. In this case, the data is the historical time-series demand of existing SKUs. PCA is used to identify and group together existing SKUs with similar demand trends. It is hypothesized that first, the new SKU will closely follow one of the trends identified by PCA. Second, the new SKU will follow the same demand trend as its most similar SKU, identified from CJA. Therefore, the subset of data from which the new SKU forecast is generated is the group of SKUs which PCA identifies as having similar demand trends and which contains the SKU that CJA determines to be the most similar to the new SKU (Figure 1)

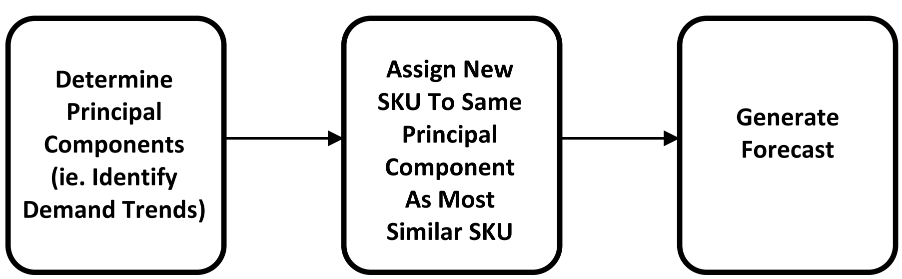

Figure 1: Schematic of Principal Component Analysis Process

2.0 Background

In anticipation of completing a thesis project, I enrolled in the Connections Program, offered through the University of Toronto, Department of Mechanical and Industrial Engineering, which offers the opportunity for students to work closely with an industry partner to fulfill the requirements of an undergrad thesis project. I approached Mr. Ed Kim of Teradata Corporation with the prospect of completing my thesis project with their support in conjunction with the support of the Ontario Center of Excellence.

Mr. Ed Kim returned my request with the opportunity to work on a project which would entail the examination of demand forecasting of new SKUs based on SKU likening concepts, an idea that was touched upon within the company by way of an internal paper, but was not yet fully realized. Further discussions with Mr. Kim expanded the scope of the project to add another level of analysis with the inclusion of PCA.

3.0 Methodology

Depicted below is a flowchart outlining the process undertaken for the completion of this thesis, with each step detailed in the sections that follow.

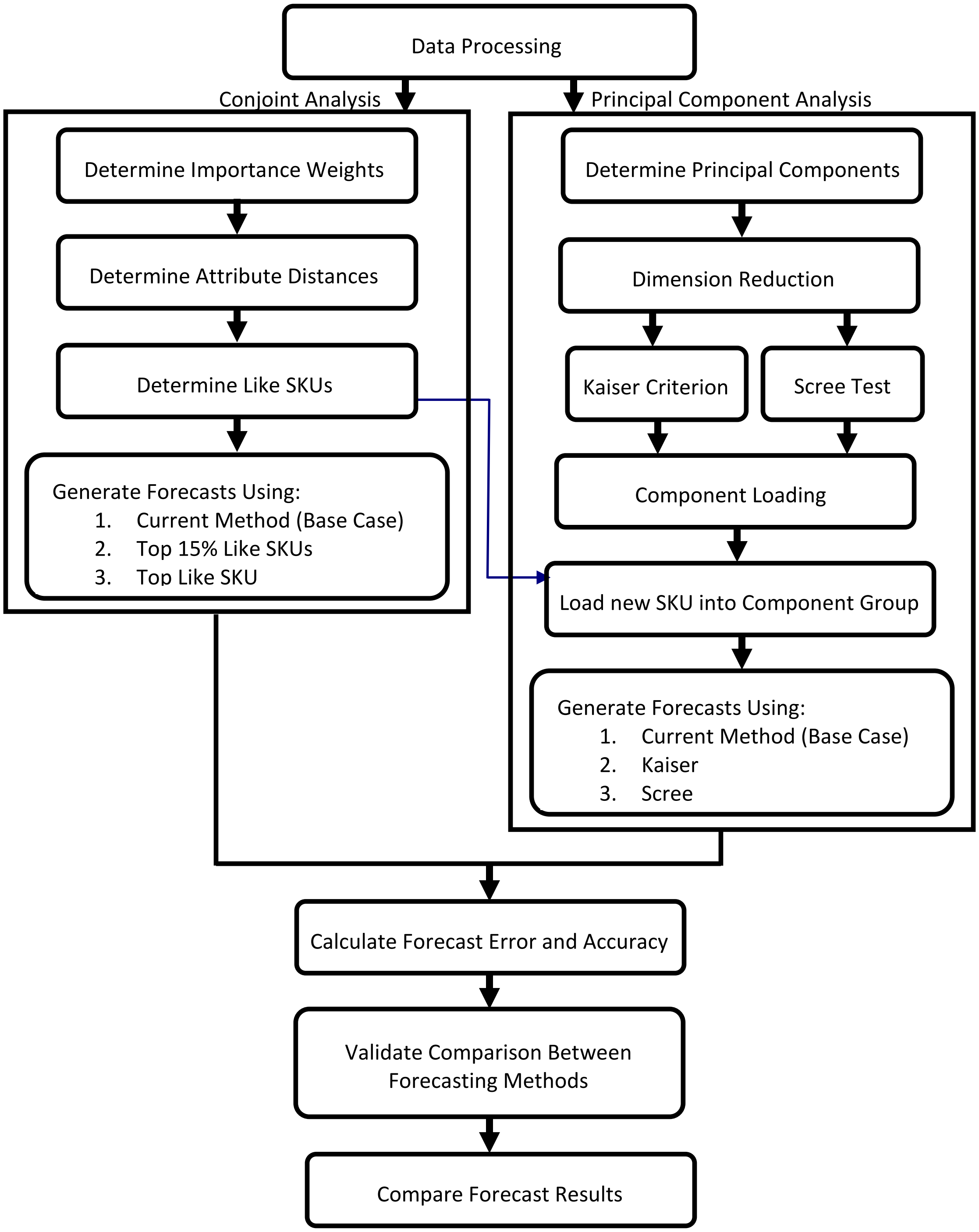

Figure 2: Thesis Experiment Process

3.1 Processing Database for Analysis

All analysis was performed on a retail database provided by the industrial partner. The data spanned two years of historical demand (2003 and 2004) and contained the following attribute information:

Data	Example
Product Number	13003
Product Code	002113008909US
Class2id	317
Class3id	4001
Class4id	0
Year-plus-Week	200301
Regular Demand	89
Regular Dollars	127.27
Year	2003
Week	01

Table 1: Information Available in the Database

Prior to the analysis, the database required reorganization in order to be useful. In the product hierarchy, the Product Number belongs to a Class4id, which belongs to a Class 3id which in turn belongs to a Class2id. Consultations with the industry partner revealed that SKUs with the same Class3id code were similar in nature and could thus be grouped together. Classifications by any other level (such as Class2 or Class4id) were not possible as no obvious pattern was revealed by their groupings. Therefore, the original data was separated into yearly files then further broken down to Year_Class3id files (for example, *2003_4001*). The final data processing action was to create a summary file, highlighting the new SKUs that appeared in 2004 in addition to any discontinued SKUs from 2003. A summary of the database is given in Table 2.

	2003	2004
Number of Class3id	37	
Number of Unique SKUs per Year	1168	1363
Number of New SKUs	n/a	210
Number of Discontinued SKUs	15	n/a

Table 2: Database Summary

For the purpose of Conjoint Analysis (described below), attributes characterizing each SKU were identified: Company, Class4id and Cost. The Company of each SKU is embedded in the first seven characters of the product code. For example, in Table 1, the company of product *13003* is *0021130*. Cost is obtained by dividing the regular dollars by the regular demand.

3.2 Conjoint Analysis

Conjoint Analysis (CJA) is a data analysis method that is used to identify similar products based on SKU attributes in a data set to select a smaller set of SKUs over which the weekly demand is averaged.

3.2.1 Background

CJA is a relatively new concept (developed in the late 1960s by marketing Professor Paul Green) that assigns values and importance weights to product attributes based on consumer feedback [5]. In our case, CJA uses multi-linear regression analysis of historical demand data to assign weighted values to each SKU attribute. Specifically, CJA returns results that illustrate which (if any) company/brand consumers have an affinity for, along with how much value they place on the product cost as well as the underlying attributes of Class4id which are not explicitly known but are still influenced by demand. Once values and importance weights are assigned to each attribute, similarity based distances between attributes of the new SKU and existing SKUs are calculated. SKUs with the shorter combined attribute distances are considered more similar than SKUs with longer distances.

To illustrate attribute distances, we consider products A, B and C with their associated attributes and attributes' importance weight listed in Table 3.

Attribute	Product A	Product B	Product C	Importance Weight		Distance A-B	Distance A-C	Distance B-C
Company	21130	13800	21130	0.46		10	0	10
Class4id	1	5	5	0.25		1	1	0
Cost	5.99	5.99	9.79	0.29		0	5	5
					Total:	11	6	15

Table 3: Attribute Distance Example

The importance weights determined from consumer response indicates that the Company/Brand attribute holds the greatest significance, followed closely by the Cost then the Class4id attribute. In this example, arbitrary weighted distance values are assigned to each attribute. For example, the distance between Companies 21130 and 12800 is 10.0. Summing the distances across all attributes reveals which products are the most similar. Considering the large significance that consumers place on Company/Brand over all other attributes, it is expected that products coming from the same Company would be deemed more similar over other pairings and thus exhibit the smallest attribute distance. This is seen with products A and C which both come from Company 21130 and are deemed the most similar by their lowest distance value of 6.0.

3.2.2 Determining Importance Weights

Although in the previous example, the attribute importance weights and values were assigned arbitrarily, the actual importance weights and values associated with SKU attributes are determined using linear regression and the formula:

$$Y = \beta_0 + x_1\beta_1 + x_{21}\beta_{21} + x_{22}\,\beta_{22} + \ldots + x_{2j}\,\beta_{2j} + x_{31}\,\beta_{31} + x_{32}\,\beta_{32} + \ldots + x_{3k}\,\beta_{3k} + \varepsilon$$

Formula 1: Calculating Importance Weights

In the above formula, Y is the demand, β_0 is the intercept term, β_1 is the significance value associated with the cost attribute, x_{1j}, x_{2j}, x_{3j} are the attributes (Cost, Class4id and Company, respectively) with j (or k) attribute levels and the remaining β's are coefficients that indicate the significance value associated with their respective attribute at the j^{th} (or k^{th}) level. For example, in Table 4, "1" and "5" are levels within the Class4id attribute. Solving the above formula for the βs returns the significance values that consumers place on each attribute level which is used to calculate the weighted importance that consumers place on the SKU attributes, Cost, Class4id and Company.

To solve the system of equations, the data must first be normalized and encoded. Each Class3id dataset of SKUs has j-levels of Class4id and k-levels of Company (which also vary across the Class3ids). In the example shown in Table 4, there are 5 levels for Company and 2 levels for Class4id. In terms of Company and Class4id, each SKU is encoded across the attribute levels such that a value of 1 represents the presence of that attribute level while a 0 represents its absence. The Cost variable, as a continuous variable, is used as is in the regression equation. Encoding all the SKUs in a Class3id dataset and making use of Excel's LINEST function solves the regression Formula 1 for the β coefficients of significance values [6].

SKU Information				Company					Class4id		
SKU	**Company**	**Class4id**	**Demand**	**0013800**	**0021130**	**0021131**	**0031000**	**0071007**	**1**	**5**	**Cost**
5459	0013800	1	4	1	0	0	0	0	1	0	6.14
131251	0021130	5	4	0	1	0	0	0	0	1	6.41

Table 4: SKU Encoding Example

Importance weights, calculated (Table 5) after the β coefficients are determined give an overall measure of how consumers value attributes relative to each other.

Attribute	Range	Importance Weight
Cost	$(\text{MAX(cost)} - \text{MIN(cost)}) * \beta_1$	$\text{range}_{cost}/\text{total\_range}$
Class4id	$\text{MAX}(\beta_{2j})-\text{MIN}(\beta_{2j})$	$\text{range}_{class4id}/\text{total\_range}$
Company	$\text{MAX}(\beta_{3k})-\text{MIN}(\beta_{3k})$	$\text{range}_{company}/\text{total\_range}$

Table 5: Importance Weight Formulas

* where $\text{total\_range} = \text{range}_{cost} + \text{range}_{class4id} + \text{range}_{company}$

3.2.2 Determining Distances

The relative distances of the attribute levels must be calculated to make attribute based comparisons between SKUs. The calculation of these distances is a two part process. The first step is to normalize the β (significance values of attribute levels) differences by dividing the difference by the total range of that particular attribute. The distance between attribute levels is then obtained by multiplying the normalized value by the attribute's importance weight (Formula 2).

$$[ABS(\beta_{attribute,j} - \beta_{attribute,j+1})/\ range_{attribute}]*\ importance\_weight_{attribute}$$

Formula 2: Calculating Company and Class4id Attribute Distances

As a continuous variable, Cost distance calculations require an extra step prior to the above calculation which entails multiplying the cost of all existing and new SKUs by the Cost significance value (β_1). The above formula, modified to calculate cost, is shown below (where m is the number of existing SKUs and n is the number of new SKUs in their respective Class3id data set):

$$[ABS(existingSKUcost_m * \beta_1 - newSKUcost_n * \beta_1)/\ range_{cost}]*\ importance\_weight_{cost}$$

Formula 3: Calculating Cost Attribute Distances

Following the above calculations, three tables listing distances between attribute levels are returned. Distances associated with Class4id and Company attribute levels form symmetric matrices (Tables 6 and 7). The values along the diagonal actually have a zero value, as there is no distance between equal attribute levels. However, keeping this in mind, the cells located along the diagonal are instead utilized to calculate the distance between the attribute at that level and an attribute level that does not appear (ie. β = 0) – thus accounting for the case in which a SKU from a new Company or new Class4id enters the market.

	Class4id_A	Class4id_B
Class4id_A	A-0_Dist	A-B_Dist
Class4id_B	A-B_Dist	B-0_Dist

Table 6: Class4id Attribute Distance Matrix

	Company_X	Company_Y	Company_Z
Company_X	X-0_Dist	X-Y_Dist	X-Z_Dist
Company_Y	X-Y_Dist	Y-0_Dist	Y-Z_Dist
Company_Z	X-Z_Dist	Y-Z_Dist	Z-0_Dist

Table 7: Company Attribute Distance Matrix

The Cost distance table forms an [n x m] table tabulating the distance between each new SKU cost against existing SKU costs (Table 8).

Existing SKU Costs	**New SKU Costs**	New_Cost_1	New_Cost_2	New_Cost_3	…
	Normalized Costs:	$New\_Cost_1 * \beta_1$	$New\_Cost_2 * \beta_1$	$New\_Cost_3 * \beta_1$	…
Ex_Cost_1	$Ex\_Cost_1 * \beta_1$	Ex_Cost_1-New_Cost_1Dist.	Ex_Cost_1-New_Cost_2Dist.	Ex_Cost_1-New_Cost_3Dist.	…
Ex_Cost_2	$Ex\_Cost_2 * \beta_1$	Ex_Cost_2-New_Cost_1Dist.	Ex_Cost_2-New_Cost_2Dist.	Ex_Cost_2-New_Cost_3Dist.	…
Ex_Cost_3	$Ex\_Cost_3 * \beta_1$	Ex_Cost_3-New_Cost_1Dist.	Ex_Cost_3-New_Cost_2Dist.	Ex_Cost_3-New_Cost_3Dist.	…
:	:	:	:	:	

Table 8: Cost Attribute Distance Matrix

3.3 Forecasting Using Conjoint Analysis Results

3.3.1 Determining Like SKUs

The first step in developing a forecast is to calculate the distances between two SKUs based on their attributes. Each new SKU is packaged with Class3id, Cost, Class4id and Company data, along with the first week it appeared in 2004. The new SKU is compared to each existing SKU of the same Class3id. From the previous distance tables, it is simply a matter of summing the total attribute distances between the new and existing SKUs. Normalizing the distances ensures that the total sum across all attributes is less than unity. Thus, a percent likeness is calculated by subtracting the total distance from 1.0. The list of comparisons between the new and existing SKUs is then sorted in descending order according to % SKU likeness.

SKU Data				
	Product No	**Company**	**Class4id**	**Cost**
New SKU	New_SKU	Company_X	Class4id_A	New_$Cost_1$
Existing SKU	Ex_SKU	Company_Y	Class4id_B	Ex_$Cost_1$
Distance Calculation				
Distance			**Total Distance**	**% Like New SKU**
Company	**Class4id**	**Cost**		
X-Y_Dist	A-B_Dist	Ex_$Cost_1$-New_$Cost_1$Dist.	(X-Y_Dist) + (A-B_Dist) + (Ex_$Cost_1$-New_$Cost_1$Dist)	1-TotalDist

Table 9: Calculating SKU Likeness

3.3.2 Three Forecasts: Base, Top 15% and Top

Three forecasts are constructed and compared using CJA results. Forecasts by current methods represents a base case, against which two other CJA based forecasts will be compared, is constructed by averaging the weekly demand of all the existing SKUs in the class, thereby creating a unique demand associated with each week of the future year (Table 10). The averaged weekly demands are then summed from the week the new SKU appears in to the end of year to create the new SKU demand forecast for the new SKU (Table 11).

Week	**SKU**	**Demand**	**SKU**	**Demand**	**SKU**	**Demand**		**Wk**	**avgDmd**
1	Ex_SKU_1	wk_1_SKU_1_dmd	Ex_SKU_2	wk_1_SKU_2_dmd	Ex_SKU_N	wk_1_SKU_N_dmd	…	**1**	$\Sigma(wk_1\_dmd)/N$
2	Ex_SKU_1	wk_2_SKU_1_dmd	Ex_SKU_2	wk_2_SKU_2_dmd	Ex_SKU_N	wk_2_SKU_N_dmd	…	**2**	$\Sigma(wk_2\_dmd)/N$
3	Ex_SKU_1	wk_3_SKU_1_dmd	Ex_SKU_2	wk_3_SKU_2_dmd	Ex_SKU_N	wk_3_SKU_N_dmd	…	**3**	$\Sigma(wk_3\_dmd)/N$
:	:	:	:	:	:	:		**:**	:
52	Ex_SKU_1	wk_{52}_SKU_1_dmd	Ex_SKU_2	wk_{52}_SKU_2_dmd	Ex_SKU_N	wk_{52}_SKU_N_dmd	…	**52**	$\Sigma(wk_{52}\_dmd)/N$
53	Ex_SKU_1	wk_{53}_SKU_1_dmd	Ex_SKU_2	wk_{53}_SKU_2_dmd	Ex_SKU_N	wk_{53}_SKU_N_dmd	…	**53**	$\Sigma(wk_{53}\_dmd)/N$

Table 10: Existing SKUs Averaged Weekly Demand

New SKU First Week	**SKU**	**Base Demand Forecast**
A	New_SKU_1	wk_A_avgDmd + wk_{A+1}_avgDmd + wk_{A+2}_avgDmd + … + wk_{53}_avgDmd
B	New_SKU_2	wk_B_avgDmd + wk_{B+1}_avgDmd + wk_{B+2}_avgDmd + … + wk_{53}_avgDmd
C	New_SKU_3	wk_C_avgDmd + wk_{C+1}_avgDmd + wk_{C+2}_avgDmd + … + wk_{53}_avgDmd
:	:	:

Table 11: Calculating Current Method Forecasts

Two forecasts for comparison are constructed. The first forecast uses the historical demand of only the Top 15% most similar SKUs from the original set. The process described above (averaging weekly demand and summing from the first week the new SKU appears) remains the same when generating the new forecast – the only exception being that the SKUs used in generating the averaged weekly demand is reduced to using only the Top 15% similar SKUs. The second forecast for comparison uses the historical demand of only the Top most similar SKU, thus simplifying the above process even further such that weekly demands do not have to be averaged as there is only one SKU in the data set from which the new SKU forecast is generated.

3.4 Principal Component Analysis

3.4.1 Background

Principal Component Analysis (PCA) is a statistical technique used to reveal patterns in complex high dimensional data sets [4]. The objective in applying PCA to this data set of existing SKUs is to group the SKUs according to similar time series demand patterns associate the new SKU with one of the groups of SKUs from which a forecast is generated. This is in contrast with the CJA method that uses the product attributes to find similar SKUs.

3.4.2 Determining Principal Components

The PCA performed for the purpose of this thesis is a four step process as follows [4]:

1. Normalize Data
2. Subtract Mean
3. Calculate the Covariance Matrix
4. Determine the Eigenvectors and Eigenvalues of the Covariance Matrix

A condensed example of four SKUs will be employed to illustrate the PCA process (Figure 3 and Table 12). A visual inspection of the plot below reveals that there exist two demand trends (one exhibited by SKUs A and B, the other by C and D) and as such, two principal components are expected to emerge.

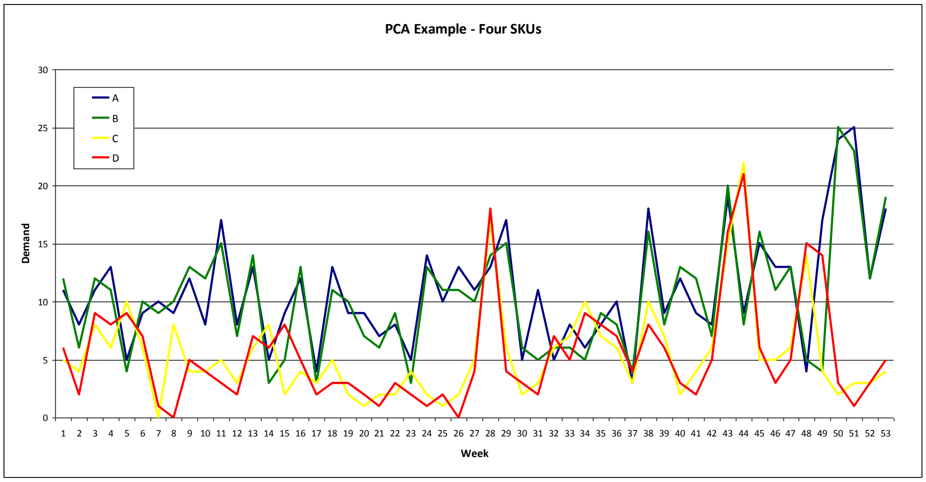

Figure 3: Four SKU PCA Example

First, the data is organized so that each SKU, forms one column in an [m x n] matrix, where m is the number of SKUs and n is the observed weekly demands (n = 53 weeks). In order to make valid comparisons across the variables, the matrix is normalized by dividing each entry by the standard deviation of the total yearly demand exhibited by the SKU (Table 12).

	Raw SKU Demand Data				Normalized Data					Normalized Data with Mean Subtracted			
week	**A**	**B**	**C**	**D**		**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
1	11	12	5	6		11/4.75 = 2.32	2.42	1.21	1.34	2.32-2.28 = 0.04	0.37	-0.12	0.13
2	8	6	4	2		8/4.75 = 1.69	1.21	0.97	0.45	1.69-2.28 = 0.60	-0.85	-0.36	-0.77
3	11	12	8	9		11/4.75 = 2.32	2.42	1.93	2.01	2.32-2.28 = 0.04	0.37	0.61	0.79
4	13	11	6	8		4/4.75 = 2.74	2.22	1.45	1.78	2.74-2.28 = 0.46	0.16	0.12	0.57
:	:	:	:	:		:	:	:	:	:	:	:	:
52	12	12	3	3		12/4.75 = 2.53	2.42	0.72	0.67	2.53-2.28 = 0.25	0.37	-0.60	-0.54
53	18	19	4	5		18/4.75 = 3.79	3.84	0.97	1.11	3.79 -2.28 = 1.51	1.78	-0.36	-0.10
stdDev:	4.75	4.95	4.14	4.49	avg:	2.28	2.06	1.32	1.21				

Table 12: Four SKU PCA Example - Sample Calculations

Second, the mean across the data set must be made to be zero, which is achieved by subtracting the column mean from each entry across all variables. This step is required for the purpose of the next step in which each dimension (each SKU) is compared to each other with respect to their means.

Third, the covariance matrix of the normalized matrix obtained in Step 2 is derived (Table 13). Covariance is a measure of how two variables vary against the mean with respect to each other. As the previous step ensured that the mean across the data set is zero, the obtained covariance matrix gives measure of trend similarity between two SKUs. Of particular note is the high level of covariance between products A-B and C-D which represents the large degree to which the two SKUs vary together, as expected.

	A	B	C	D
A	1	0.8731	-0.0319	0.0189
B	0.8731	1	0.0123	-0.028
C	-0.0319	0.0123	1	0.8549
D	0.0189	-0.028	0.8549	1

Table 13: Covariance Matrix

The final step in PCA is to find the eigenvectors and eigenvalues of the covariance matrix obtained previously (Table 14). The eigenvectors of the covariance matrix represent a rotation of the original data onto a set of axes that better highlight trends hidden in the data. The corresponding eigenvalues represent the amount of variance explained by the eigenvector. The objective is to identify the eigenvectors in which variance (that is, the signals or different trends present in the data) is the most significant, or highest. As such, the eigenvector with the corresponding highest eigenvalue is termed the first principal component as it is the principal component which explains for the majority of the variance found in the data.

	Principal Component 1	Principal Component 2	Principal Component 3	Principal Component 4	Eigenvalue
A	-0.61966	0.34074	-0.44648	0.548244	1.881008
B	-0.6201	0.339694	0.449873	-0.54562	1.846983
C	0.34208	0.618821	0.547313	0.447762	0.182412
D	0.338357	0.620939	-0.54655	-0.44859	0.089597

Table 14: Eigenvectors and Eigenvalues to Determine Principal Components

3.4.2 Dimension Reduction

Following PCA, the matrix comprised of the resulting eigenvectors is of same dimension as the initial data set (that is, the number of existing SKUs) with each eigenvector accounting for a portion of the overall variance present in the data. However, the amount of variance represented by each component diminishes as their corresponding eigenvalues decrease. Therefore, it is possible to eliminate components of low eigenvalues without significant loss in data.

Currently there are two popular criterions used to determine the number of components to retain: the Kaiser criterion (also known as Rule of One) and the Scree test [7]. This thesis employs both criteria for comparison purposes. According to the Kaiser criterion, all components with eigenvalues less than one may be discarded (Table 15). The Scree test (Figure 4), on the other hand plots the decreasing eigenvalues in a simple line plot and makes note of where the slope of the curve makes a sharp change (most notably in the "knee" of the curve). The components to retain, then, are to the left of the knee. For the four SKU example used above, both the Kaiser and SCREE Test reveal that only two Principal Components should be retained.

Kaiser Criterion
EigenValues
1.881008
1.846983
0.182412
0.089597

Table 15: Kaiser Criterion

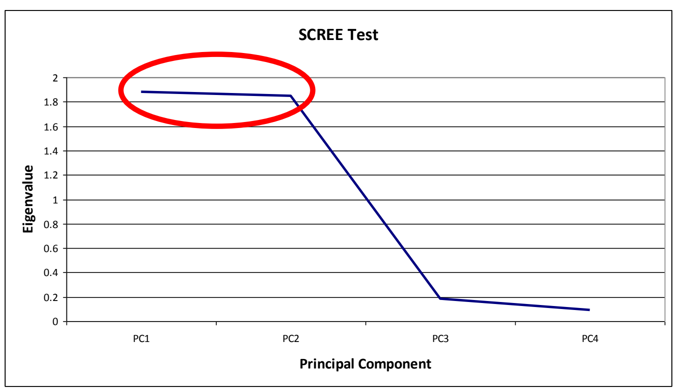

Figure 4: SCREE test

3.4.3 Component Loading and Significance

Once the number of principal components to retain is determined, it is necessary to load each existing SKU into a component group. This is accomplished by rotating the components (performed in MATLAB using the *rotatefactors* function [8]) so as to maximize the variance between them and highlight which component best explains the trend associated with a given SKU. Each component contains data associated with each SKU and each SKU is described by a sum of all the components, yet only one component (the component with the maximum absolute value for that SKU) best describes each SKU. As Table 16 highlights, SKUs A and B are both loaded onto Principal Component 1 and SKUs C and D are loaded onto Principal Component 2.

Rotated Factors		
	PC1	PC2
A	-0.69808	0.325991
B	-0.71591	-0.30833
C	-0.0088	0.886908
D	0.008831	0.886474

Table 16: Rotated Factors and Component Loading

At this stage, statistical analysis is performed on the component groups to ensure that the PCA had succeeded in grouping the SKUS according to trend similarity. The theory is that grouping the SKUs according to their time series demand patterns results in noise reduction when comparing the noise present in the data set as a whole versus noise present within each component group.

A metric for determining the noise present in data is to calculate the standard deviation of errors in the data. The errors are calculated as deviation of the actual signal from the trend's average signal. Organizing a table such that each row consists of the weekly normalized demand for one SKU and sorting the table by Component groupings, weekly averages are taken across the data set as a whole, then across each component group. These averages represent the average signal present within the whole data set and within each Component group. The deviation (or error) from each SKU to its Component group's average signal and the overall average signal were taken (Table 17). Finally, the standard deviations associated with the signal errors within factor groups and with signal errors within the whole group are taken and compared.

Normalized Data											
		1	2	3	4	5	…	51	52	53	
PC1	A	2.317546	1.685488	2.317546	2.738919	1.05343	…	5.267151	2.528233	3.792349	
	B	2.424258	1.212129	2.424258	2.222237	0.808086	…	4.646495	2.424258	3.838409	
PC2	C	1.206386	0.965109	1.930218	1.447663	2.412772	…	0.723832	0.723832	0.965109	
	D	1.337136	0.445712	2.005704	1.782848	2.005704	…	0.222856	0.668568	1.11428	
Average Signals											
All		1.821332	1.07711	2.169432	2.047917	1.569998	…	2.715084	1.586223	2.427537	
PC1		2.370902	1.448809	2.370902	2.480578	0.930758	…	4.956823	2.476245	3.815379	
PC2		1.271761	0.70541	1.967961	1.615256	2.209238	…	0.473344	0.6962	1.039695	
Deviation From Average Signal (Component Groups)											
A		2.32-2.37 = -0.053	0.23668	-0.05336	0.258341	0.122672	…	0.310328	0.051987	-0.02303	STD DEV 0.274
B		2.42-2.37 = 0.053	-0.23668	0.053356	-0.25834	-0.12267	…	-0.31033	-0.05199	0.02303	
C		1.21-1.27 = -0.065	0.259698	-0.03774	-0.16759	0.203534	…	0.250488	0.027632	-0.07459	
D		1.34-1.27 = 0.065	-0.2597	0.037743	0.167593	-0.20353	…	-0.25049	-0.02763	0.074586	
Deviation From Average Signal (All)											
A		2.32-1.82 = 0.496	0.608379	0.148115	0.691002	-0.51657	…	2.552068	0.94201	1.364812	STD DEV 0.862
B		2.42 -1.82 = 0.603	0.13502	0.254827	0.17432	-0.76191	…	1.931412	0.838036	1.410872	
C		1.21 – 1.82-0.615	-0.112	-0.23921	-0.60025	0.842774	…	-1.99125	-0.86239	-1.46243	
D		1.34 – 1.82 = -0.484	-0.6314	-0.16373	-0.26507	0.435706	…	-2.49223	-0.91765	-1.31326	

Table 17: Component Loading Significance with Some Sample Calculations

The Component groupings are a success if the standard deviation associated with the Component groups (and subsequently the noise) is less when compared to the standard deviation of errors associated with the whole data set. In the above example, the resulting standard deviations indicate that the component groupings are indeed a success.

3.5 Forecasting Using PCA

Assuming the previous section produced favorable results and the SKUs are successfully grouped by trend, the next step is to load each new SKU into a Component group. After loading the new SKU into a Component group that is believed will exhibit a similar demand trend, a forecast for the new SKU is developed using the historical demand of the SKUs within the same Component group.

CJA results are used to determine which Component group into which the new SKU will belong. New SKUs are loaded into the same Component group as the SKU that CJA identified as being the most similar. The objective is that the CJA results will be enhanced by the PCA results, which reduced noise present in the data, something that was not taken into consideration in the pure CJA forecasts.

Forecasts are generated exactly like previous forecast generation, but applied to a different data set. First, the weekly demand of the SKUs within each component group is averaged. The new SKU forecast is then generated by summing the averaged weekly demand of the component group it was loaded into, beginning from the week in which it first appeared to the end of year.

3.6 Forecast Error, Accuracy and Statistical Significance

Forecast results are compared and analyzed according to the forecasts' error and accuracy, in conjunction with statistical verification that a valid comparison can be made among the different forecasting methods employed.

Individually, forecast error and accuracy for each new SKU are defined as follows:

$$\begin{aligned} |\text{error}| &= |\text{actual} - \text{forecast}| \\ \text{error} &= |\text{ error }| / \text{actual} \\ \text{accuracy} &= \text{MAX}[1\text{-error}, 0] \end{aligned}$$

Across a group of SKUs, the overall error is given by:

$$\text{error} = \Sigma\,|\text{ error }|\,/\,\Sigma\text{ actual}$$

Prior to making any comparisons, it first must be verified that valid comparisons can be made among the resulting forecasts. This is accomplished with the use of hypothesis testing which tests the underlying population from which the data emerges. The test concludes that with a stipulated confidence, all forecasts and their resulting errors are from the same underlying population and thus meaningful conclusions can be inferred from the forecast results when comparisons are made. To wit, the following hypothesis is tested:

$$H_o: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Where μ_1 is the mean of the base forecast error group and μ_2 is the mean of one of the proposed forecast error groups. The confidence level is set at 95%, thus, a resulting p-value of less than 0.05 indicates, with 95% confidence, it can be stated that the two forecast groups in question have the same underlying population and thus meaningful conclusions can be made when making comparisons between forecasting methods when testing them against the base case.

4.0 Results and Discussion

4.1 Data

Forecasting and subsequent comparisons and analyses were done within each Class3id attribute group. As such, it was imperative that a sufficient sample size of new SKUs was present in each of the analyzed groups in order for subsequent error analysis to hold significance. Therefore, the Class3id groups were sorted according to the number of new SKUs that appeared in the second year, and the top five classes containing the most new SKUs were taken for analysis. New SKUs present in the top five Class3id groups accounted for 129 out of 210 new SKUs that appeared in the second year. A summary of the Class3id groups used and the number of new SKUs present in each group is tabulated in Table 18.

Class3id	No. of New SKUs
4801	41
4201	35
4804	23
4802	17
4820	13
Total:	129

Table 18: Data Summary

For the purpose of conciseness, the following sections will detail results pertaining only to class3ID **4802** (arbitrarily chosen), with overall results detailed where necessary.

4.2 Conjoint Analysis

Each Class3id analysis contained the following information (performed in Excel):

- a master sheet listing all the weekly data for products in that class
- a “weights” sheet determining the values and importance weights of each attribute and attribute level
- a distance sheet calculating the distances between attribute levels
- sheets for each new SKU containing distance calculations between the new SKU and the existing SKUs as well as the weekly data for the SKUs identified as being similar
- an error report tabulating and comparing actual demands to forecasts generated using current methods (the “base” forecast) and like-SKU forecasted demands.

First, consumer demand was used to determine significance values for each attribute level. From Table 19, it is noted that consumers place large value on the Cost attribute, Class4id **1** attribute level and on Companies/Brands **0013800** and **003100**.

		Class4id		Company					
demand	**cost**	**5**	**1**	**0071007**	**0031000**	**0021131**	**0021130**	**0013800**	**intercept**
coefs	0.9	0	7.83	0	-9.74495	-10.622	0	-14.644	2.96444
stderr	0.6	0	3.72	0	4.70738	6.01775	0	5.53732	3.9732
r2,sev	0.2	5.99							
F, df	1.6	28							
ssreg,ssregid	290	1004.6							
t-value	1.5	#DIV/0!	2.1	#DIV/0!	-2.07014	-1.7651	#DIV/0!	-2.6445	0.74611
p-value	0.2	#DIV/0!	0.04	#DIV/0!	0.04778	0.08846	#DIV/0!	0.01326	0.46182

Table 19: Attribute Level Significance Values for 2003_Class3id_4802

Second, equations found in Table 5 were used to calculate the attributes importance weights (iw):

attribute	**range**	**iw**
cost	9.123017	0.288703
class4id	7.8335	0.247895
company	14.64351	0.463402
Σ	31.60003	1

Table 20: Attribute Importance Weights

From Table 20, it is observed that for SKUs in Class3id 4802, upon purchase, consumers place almost equal weight on the Cost and Class4id attributes of the product and greater value to its Company/Brand.

Finally, the above importance weights were applied to a table of attributes to determine the relative distances of the attribute levels, as illustrated in Table 21 for the Class4id and Company attributes, and partially in Table 22 for the Cost attribute. The highlighted values are for reference in a later sample calculation for determining attribute based SKU similarities.

		Class4id		Company				
		5	1	0071007	0031000	0021131	0021130	0013800
Class4id	5	0	0.247895					
	1	0.247895	0.247895					
Company	0071007			0	0.308384	0.336129	0	0.463402
	0031000			0.308384	0.308384	0.027745	0.308384	0.155018
	0021131			0.336129	0.027745	0.336129	0.336129	0.127273
	0021130			0	0.308384	0.336129	0	0.463402
	0013800			0.463402	0.155018	0.127273	0.463402	0.463402

Table 21: Distances Between Class4id and Company Attribute Levels

Existing SKU Costs		New SKU costs						
		3.7	4.4	4.43	4.46	4.47	4.69	…
		3.442652	4.093965	4.121878	4.149792	4.159096	4.363794	…
2.91	2.710369	0.023174	0.043785	0.044668	0.045551	0.045846	0.052324	…
3.15	2.929748	0.016231	0.036842	0.037726	0.038609	0.038903	0.045381	…
3.15	2.932779	0.016135	0.036746	0.03763	0.038513	0.038807	0.045285	…
3.15	2.933484	0.016113	0.036724	0.037607	0.038491	0.038785	0.045263	…
3.16	2.936741	0.01601	0.036621	0.037504	0.038388	0.038682	0.04516	…
3.30	3.068001	0.011856	0.032467	0.033351	0.034234	0.034528	0.041006	…
3.35	3.117936	0.010276	0.030887	0.03177	0.032654	0.032948	0.039426	…
3.47	3.231315	0.006688	0.027299	0.028182	0.029066	0.02936	0.035838	…
3.48	3.240869	0.006386	0.026997	0.02788	0.028763	0.029058	0.035536	
:	:	:	:	:	:	:	:	

Table 22: Distances Between Existing and New SKU Cost Attribute Levels

Once all the weights and distances were calculated, data from the new SKU were brought in for comparison which is best illustrated with an example. New SKU 13281 has the following information for its introductory week:

product ID	productcode	class4id	yearplusweekno	regulardemand	regulardollars	cost
13281	002113010706US	5	200408	3	9	$3.70

Table 23: Product Information for New SKU

For SKU likening purposes, the cost of the new SKU was taken to be its average cost over the rest of the year. The average cost of SKU 13281 was $3.70.

The new SKU attributes were compared with the existing SKU attributes in the Class3id as shown in Table 24 where the distance values were taken from distance tables shown in Tables

21 and 22 above. Comparisons between the new SKU and the remaining existing SKUs were made in the same manner and the table was then sorted according to the % Like New SKU column, in decreasing order.

	Product ID	Product Code	Company	Class4id	Cost	Distance			Total Distance	%Like New SKU
New SKU	13281	002113010706US	0021130	5	3.70	Company	Class4id	Cost		
Existing SKU_1	13313	002113010706US	0021130	5	3.48	0	0	0.00639	0.00639	0.99361
Existing SKU_2	65953	007100701001US	0071007	1	3.15	0	0.24790	0.01623	0.263905	0.73609

Table 24: Distance Calculation Example

4.3 Forecasting Using CJA

The above procedure was followed for the remaining 16 new SKUs of 2004 and an error report (Table 25 and Figure 5) was generated tabulating the error between existing forecasting methods (the base case) and SKU-likening forecasting methods. Actual demand of the new SKU was taken from the 2004 database and forecasts were generated using the formulas and methods detailed in 3.3.2

wk	prodID	Actual Rest of Year Demand	Base Demand	Top Like SKU Demand	Top 15% Like SKUs Demand	Base Demand |Err|	Top Like SKU |Err|	Top 15% Like SKUs |Err|	Base Demand Err or	Top Like SKU Err or	Top 15% Like SKUs Error	Base Demand Accuracy	Top Like SKU Accuracy	Top 15% Like SKUs Accuracy
6	p5452	82	392	157	268	310	75	186	3.78	0.91	2.27	0.00%	8.54%	0.00%
13	p5613	248	322	80	212	74	168	36	0.30	0.68	0.15	70.01%	32.26%	85.42%
12	p5615	39	330	82	219	291	43	180	7.47	1.10	4.62	0.00%	0.00%	0.00%
8	p13281	168	373	190	320	205	22	152	1.22	0.13	0.90	0.00%	86.90%	9.71%
42	p13348	17	97	59	48	80	42	31	4.71	2.47	1.85	0.00%	0.00%	0.00%
40	p13349	67	115	59	84	48	8	17	0.71	0.12	0.25	28.86%	88.06%	75.35%
41	p13350	98	106	77	76	8	21	22	0.08	0.21	0.23	91.64%	78.57%	77.26%
41	p13351	86	106	77	76	20	9	10	0.23	0.10	0.12	76.52%	89.53%	88.04%
40	p13393	55	115	67	84	60	12	29	1.08	0.22	0.52	0.00%	78.18%	48.15%
40	p13400	102	115	67	84	13	35	18	0.12	0.34	0.18	87.58%	65.69%	81.88%
41	p13401	55	106	64	76	51	9	21	0.93	0.16	0.38	6.92%	83.64%	62.33%
50	p13402	5	35	22	17	30	17	12	5.94	3.40	2.47	0.00%	0.00%	0.00%
41	p13403	43	106	60	52	63	17	9	1.47	0.40	0.22	0.00%	60.47%	78.22%
22	p27757	141	248	195	217	107	54	76	0.76	0.38	0.54	24.24%	61.70%	46.13%
28	p27758	135	205	179	188	70	44	53	0.52	0.33	0.39	48.41%	67.41%	60.85%
22	p27759	128	248	195	217	120	67	89	0.94	0.52	0.69	6.39%	47.66%	30.51%
22	p27762	158	248	195	217	90	37	59	0.57	0.23	0.37	43.15%	76.58%	62.69%
	Σ	1627	3267	1825	2453	1640	680	1001	1.0078	0.42	0.61499	0	58.21%	38.50%

Table 25: CJA Based Class3id_4802 Error Report

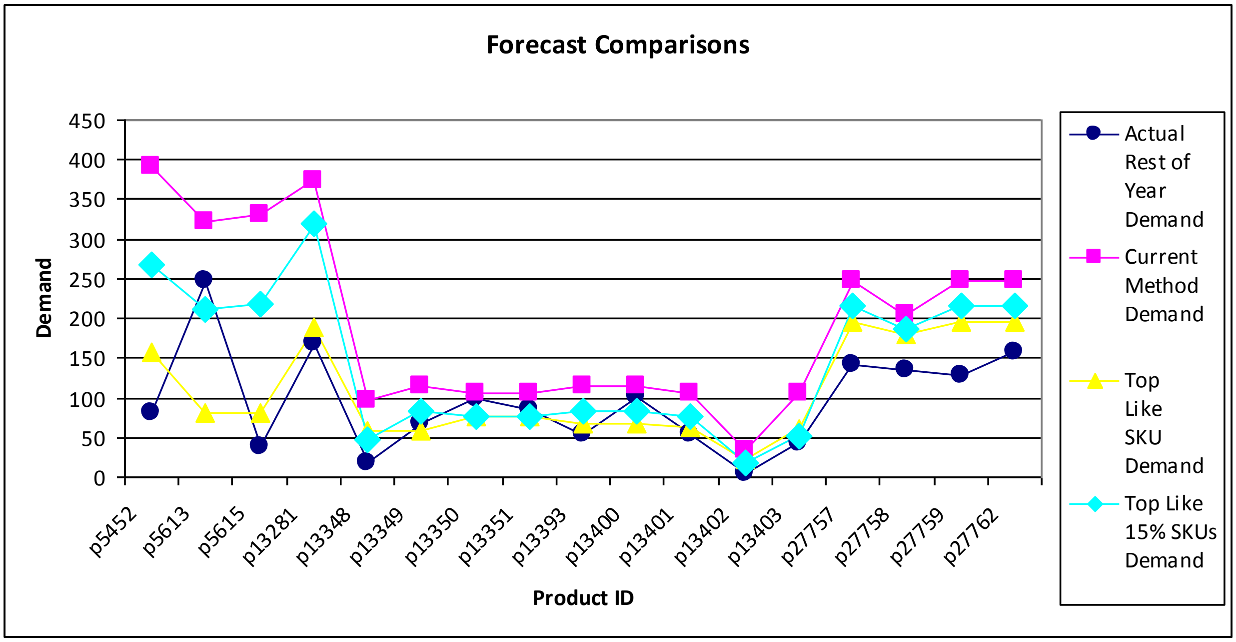

Figure 5: Forecasts Developed Using Current Methods and CJA Based Methods

Although the results look promising, prior to making any comparisons, the validity of making comparisons between the forecasting methods must be verified. That is, it must be stated with at least 95% confidence that the errors resulting from these forecasts belong to the same larger population. Two comparisons are made. The first is between the base error group and most similar SKU error group, which returned a p-value of 0.018. The second comparison was between the base error group and top 15% similar SKUs which returned a p-value of 0.000. Therefore, based on these results, it can be stated with 95% confidence that these two data sets are part of the same population and conclusions made from comparisons between the groups are valid.

As the Table 25 and Figure 5 detail, for this particular class, current forecasting methods failed to generate a forecast of any accuracy. In comparison between forecasts using the top 15% similar SKUs versus using only the most similar SKU, the most similar SKU forecast outperformed the former by nearly 20%.

Overall, CJA based forecasts generated for the 129 new SKUs, produced forecasts of improved accuracy over current method forecasts. The forecast analysis summary is tabulated below (Table 26):

Σ (Actual Rest of Year Demand)	Σ (Current Method Demand)	Σ (Top Like SKU Demand)	Σ (Top Like 15% SKUs Demand)	Σ (Base Demand |Err|)	Σ (Top Like SKU |Err|)	Σ (Top Like 15% SKUs |Err|)	Base Demand Err %	Top Like SKU Err %	Top Like 15% SKUs Err %	Base Demand Accuracy	Top Like SKU Accuracy	Top Like 15% SKUs Accuracy
21181	44930	35187	39209	24993	18436	18871	1.18	0.87	0.89	0	12.96%	10.91%

Table 26: CJA Based Forecast Summary

The results were first validated with the hypothesis test comparing the current forecasting method error group against the CJA based forecast error groups with p-values of 0.00037 and 1.86E-09, for the most similar SKU and top 15% similar SKUs, respectively.

The overall results show that of current methods produced forecasts of zero accuracy. Of the two CJA based forecasts, the forecast generated using only the historical demand of the most similar SKU produced better accuracies versus using historical demand of the top 15% similar SKUs.

4.4 Principal Component Analysis

Briefly restated, the ultimate objective of performing PCA on the data of existing SKUs within each Class3id, is to reveal time series demand trends among the existing SKUs, and group them according to the identified trends. For example, the graph below (Figure 6) depicts the normalized weekly demands of all SKUs appearing in Class3id 4802, which does not exhibit any single obvious demand trend, but rather many demand trends grouped together. The objective of PCA is to identify and isolate the trends, such that SKUs can be grouped together as seen in Figures 7 and 8.

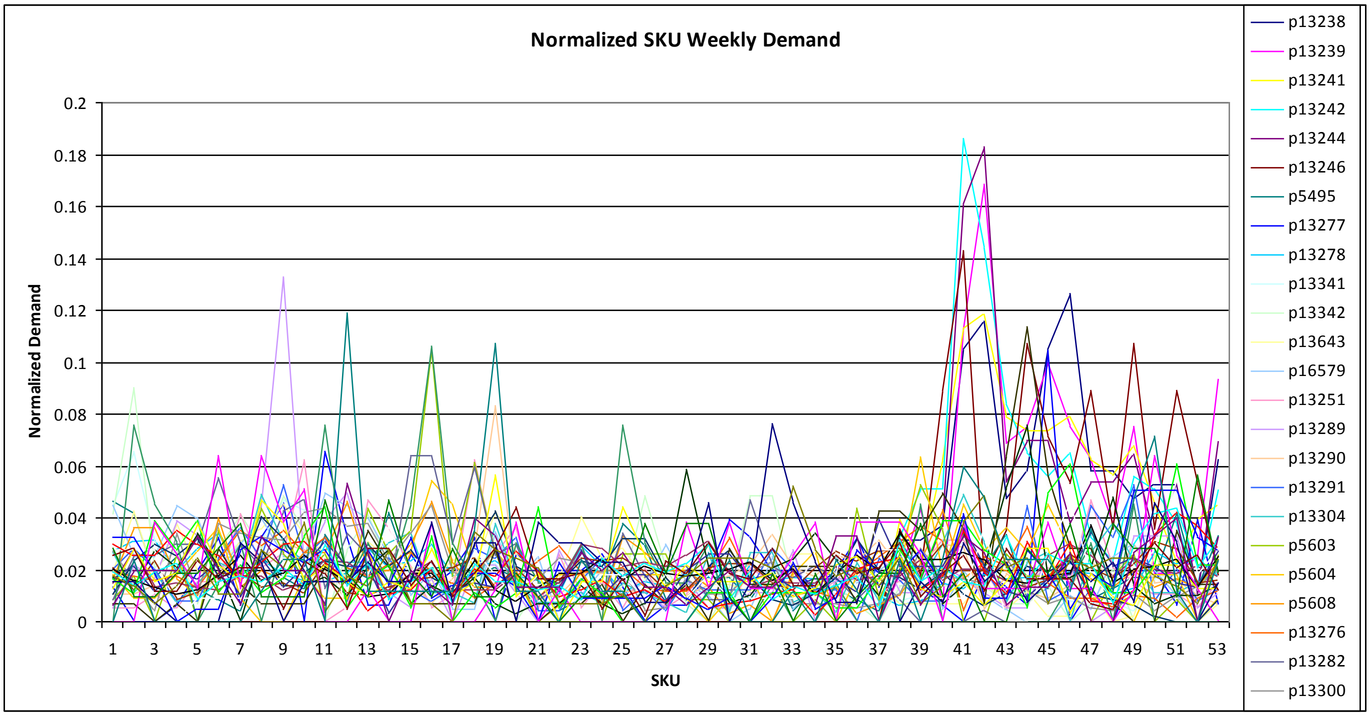

Figure 6: PCA – Normalized SKU Weekly Demand

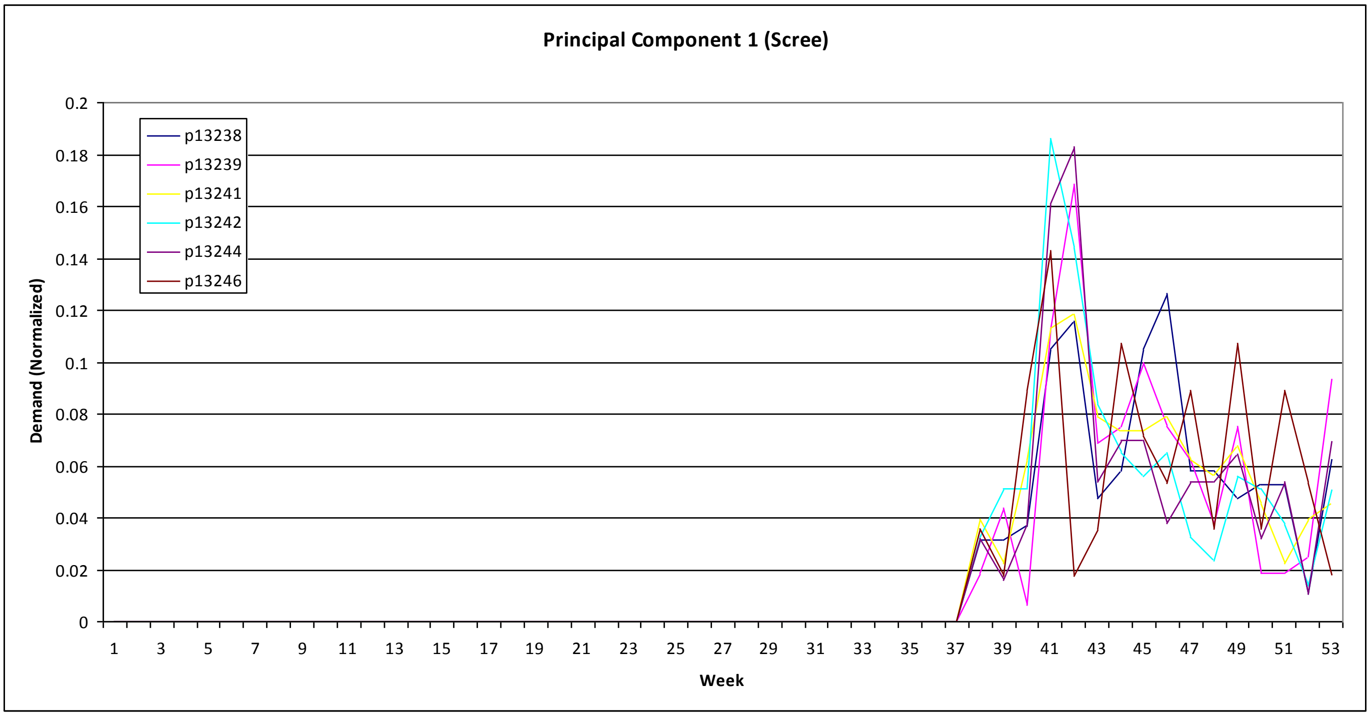

Figure 7: PCA – Scree Principal Component Group 1

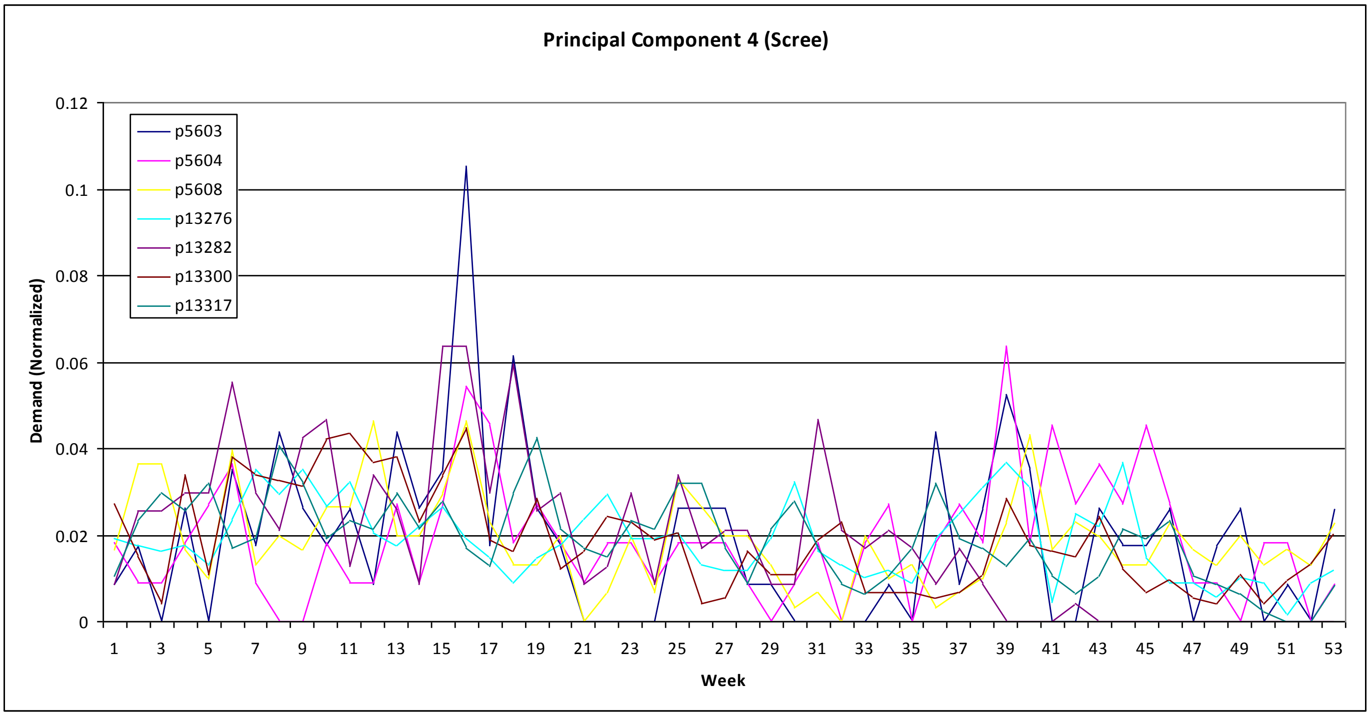

Figure 8: Scree – Principal Component Group 4

Class3id 4802 was comprised of 45 SKUs. Performing PCA on the data set of 45 SKUs returned a matrix of 45 components and their associated eigenvalues. However, from the 45 components, only the number of components as determined by the Kaiser criterion and Scree test would remain. Of the 45 components, only the top 14 had eigenvalues greater than one, thus according to the Kaiser criterion, only 14 components would be retained. Alternatively, the Scree test (shown in Figure 9) reveals that only seven components should be retained, as evidence by the knee in the curve located approximately at the seventh component where the curve inexplicably begins to level off.

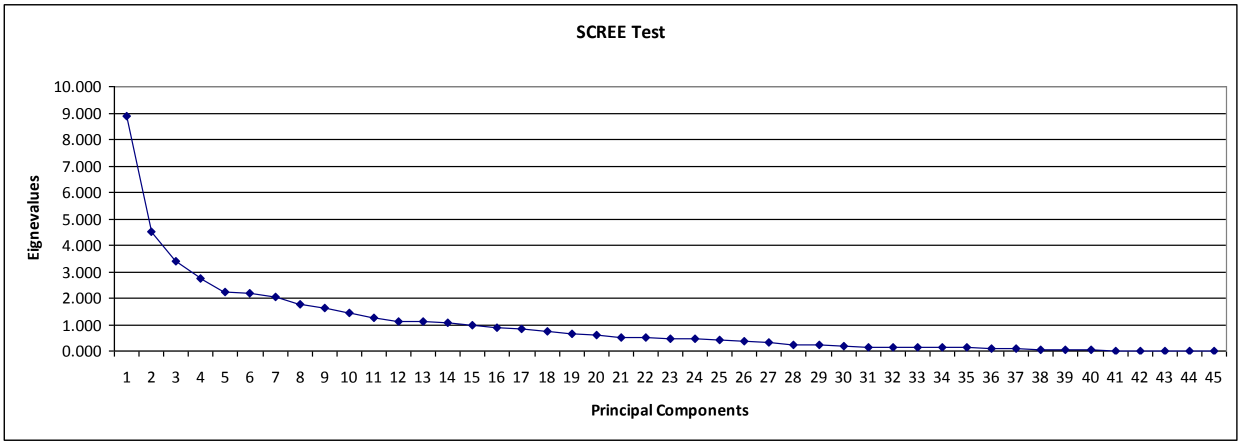

Figure 9: SCREE Test

The next steps were to load each SKU into a principal component group and verify the success of the groupings. Table 27 lists all the SKUs and the component groups into which they were loaded. Each SKU is loaded into two groups, one according to the Kaiser criterion and one according to the Scree test. The weekly demands of the products shaded in yellow and grey are shown in above Figures 7 and 8, respectively. Visually, the graphs show a clear trend among these groups of SKUs after they have been loaded onto a component. The next step is to statistically verify the success.

prodNo	PC(Kaiser)	PC(Scree)	prodNo	PC(Kaiser)	PC(Scree)	prodNo	PC(Kaiser)	PC(Scree)
5495	10	2	13251	7	3	13332	10	6
5600	11	6	13276	6	4	13334	9	7
5601	5	7	13277	8	2	13335	9	7
5602	10	6	13278	6	2	13341	13	2
5603	4	4	13282	7	4	13342	13	2
5604	4	4	13289	8	3	13643	13	2
5605	3	5	13290	12	3	16570	2	6
5606	11	6	13291	7	3	16579	5	2
5608	14	4	13300	4	4	27707	6	5
13238	1	1	13304	5	3	27740	14	6
13239	1	1	13310	14	7	27747	12	5
13241	1	1	13313	2	6	65953	6	5
13242	1	1	13317	12	4	65955	3	5
13244	1	1	13318	12	6	65956	10	5
13246	2	1	13323	7	7	147612	6	5

Table 27: Component Loading of SKUs

The success of the groupings is verified by calculating the amount of noise present in the data set as a whole and comparing it to the noise present in the data set after it has been grouped into components. Standard deviation as a measurement of noise, will verify the success of the loadings. The table below shows the standard deviation results over all the analyzed classes. As the values indicate, there is a definite noise reduction as the data goes from being analyzed as a whole, to being analyzed according to the Scree groupings and finally, the most reduction in noise occurs when the SKUs are loaded into components as determined by the Kaiser criterion.

Class:	**4201**	**4801**	**4802**	**4804**	**4820**
Number of existing SKUs	177	223	45	80	66
Number of new SKUs	35	41	17	23	13
Number of components (Kaiser)	33	33	14	25	19
Number of components (SCREE)	7	7	7	8	5
stdDev (Kaiser)	0.0136	0.013	0.0103	0.0133	0.0094
stdDev (Scree)	0.0151	0.0203	0.0118	0.0163	0.0115
stdDev (Ungrouped)	0.017	0.0263	0.0173	0.0177	0.014

Table 28: Significance of Component Loading

Next, forecasts will reveal whether or not forecasts based on the Kaiser criterion will outperform forecasts based on the Scree test, as is hypothesized from the above results.

4.5 Forecasting Using PCA

Forecasts generated using PCA first involved loading each new SKU into one of the component groups which made use of the CJA results. Simply, the new SKUs were loaded into the component group as its most similar SKU determined from CJA. Once the new SKUs were loaded into a component group, a forecast was generated from the summed averaged weekly demands exhibited by the SKUs in its component group. The results for class 4802 are tabulated below:

wk	SKU	Top Like SKU	KAI	SCR	Actual Rest of Year Dmd	Base Dmd	KAI Dmd	SCR Dmd	Base Dmd |Err|	KAI |Err|	SCREE |Err|	Base Dmd Error	KAI Error	SCREE Error	Base Dmd Acc	KAI Acc	SCREE Acc
6	p5452	5606	11	6	82	375	128	294	293	46	212	3.57	0.56	2.58	0.00%	43.90%	0.00%
13	p5613	5601	5	7	248	310	130	120	62	118	128	0.25	0.47	0.52	75.04%	52.55%	48.31%
12	p5615	5601	5	7	39	318	136	125	279	97	86	7.14	2.49	2.21	0.00%	0.00%	0.00%
8	p13281	13313	2	6	168	357	251	281	189	83	113	1.13	0.49	0.67	0.00%	50.60%	32.59%
42	p13348	13318	12	6	17	96	86	88	79	69	71	4.65	4.06	4.19	0.00%	0.00%	0.00%
40	p13349	13313	2	6	67	115	110	103	48	43	36	0.71	0.65	0.54	28.69%	35.32%	46.27%
41	p13350	13313	2	6	98	107	104	97	9	6	1	0.09	0.06	0.01	91.04%	94.22%	99.11%
41	p13351	13313	2	6	86	107	104	97	21	18	11	0.24	0.21	0.13	75.84%	79.46%	87.06%
40	p13393	13318	12	6	55	115	109	103	60	54	48	1.09	0.99	0.87	0.00%	1.36%	12.73%
40	p13400	13318	12	6	102	115	109	103	13	7	1	0.13	0.07	0.01	87.47%	92.89%	99.02%
41	p13401	13318	12	6	55	107	97	97	52	42	42	0.94	0.76	0.77	5.86%	23.64%	23.41%
50	p13402	13318	12	6	5	32	21	40	27	16	35	5.43	3.20	7.08	0.00%	0.00%	0.00%
41	p13403	13317	12	4	43	107	97	58	64	54	15	1.48	1.26	0.34	0.00%	0.00%	65.78%
22	p27757	27707	6	5	141	240	498	319	99	357	178	0.70	2.53	1.27	30.02%	0.00%	0.00%
28	p27758	27707	6	5	135	199	408	265	64	273	130	0.47	2.02	0.96	52.86%	0.00%	3.60%
22	p27759	27707	6	5	128	240	498	319	112	370	191	0.87	2.89	1.50	12.76%	0.00%	0.00%
22	p27762	27707	6	5	158	240	498	319	82	340	161	0.52	2.15	1.02	48.31%	0.00%	0.00%
				Σ	1627	3177	3385	2830	1550	1993	1461	0.95	1.23	0.90	4.73%	0.00%	10.19%

Table 29: PCA Based Class3id_4802 Error Report

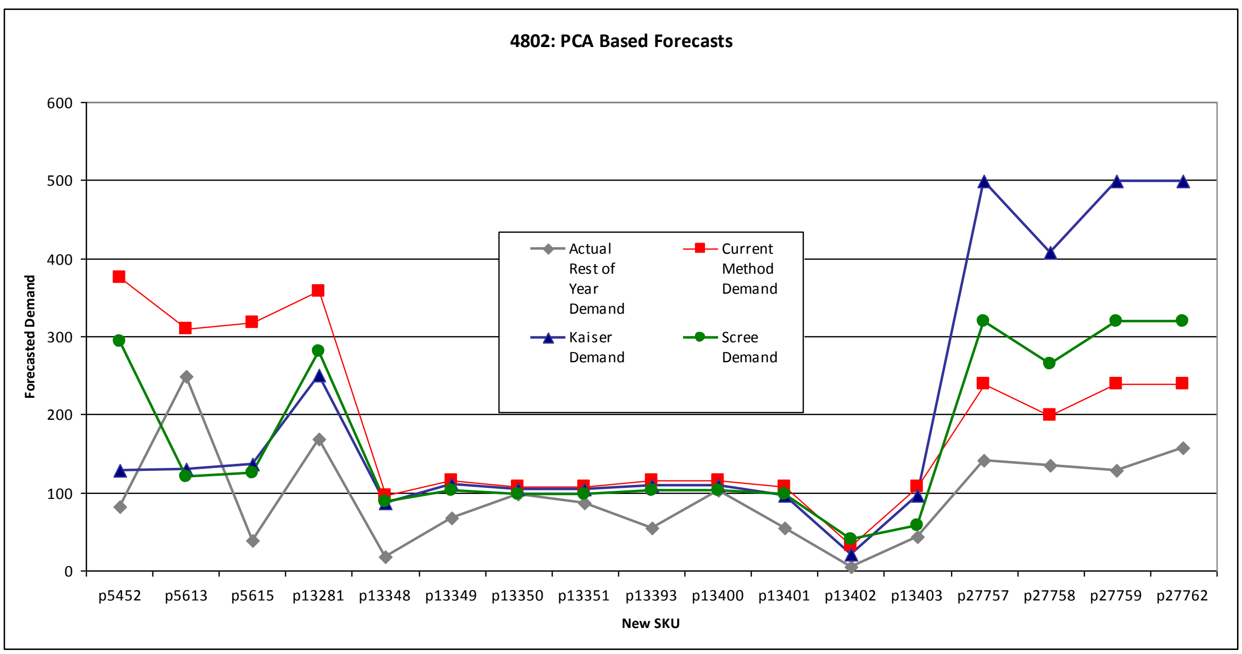

Figure 10: Forecasts Developed Using Current Methods and PCA Based Methods

The results from the PCA based forecasts are not promising. Forecasts generated from the SKUs loaded into Kaiser based component groups produced zero accuracy overall. The Scree based component groups faired slightly better with 10.19% accuracy, which was still an improvement over base demand accuracy of 4.73%. Tabulating results spanning all five classes, however, returned somewhat better results. The results showed that the base demand forecasted with zero accuracy, the Scree based component groups produced forecasts of 9.95% accuracy while the Kaiser based component groups produced the best overall accuracy of 13.46%.

The above comparisons were validated with the hypothesis test comparing the current forecasting method error group against the PCA based forecast error groups. Resulting p-values of 0.015 and 0.034, for the Kaiser and Scree based forecasts, respectively, validated the comparisons within the stipulated 95% confidence.

A deeper analysis into the results revealed that in most cases, the generated forecasts were heavily biased to one side of the actual demand – that is, all the forecasts were biased to generate much greater forecasts than the actual demand. Additionally, comparing the average demand within each component group against the average demand of the new SKUs' most similar SKU shows that the most similar SKU often had a significantly lower yearly average (Table 30).

wk	prodID	Top Like SKU	Top SKU avgDmd	KAI PC	KAI avg fcst	SCR PC	SCR avg fcst
6	p5452	5606	4	11	3	6	6
13	p5613	5601	2	5	4	7	3
12	p5615	5601	2	5	4	7	3
8	p13281	13313	4	2	5	6	6
42	p13348	13318	6	12	9	6	6
40	p13349	13313	4	2	5	6	6
41	p13350	13313	4	2	5	6	6
41	p13351	13313	4	2	5	6	6
40	p13393	13318	6	12	9	6	6
40	p13400	13318	6	12	9	6	6
41	p13401	13318	6	12	9	6	6
50	p13402	13318	6	12	9	6	6
41	p13403	13317	9	12	9	4	7
22	p27757	27707	6	6	16	5	9
28	p27758	27707	6	6	16	5	9
22	p27759	27707	6	6	16	5	9
22	p27762	27707	6	6	16	5	9

Table 30: Top Like SKU Yearly Average versus Component Group Yearly Average

This indicated that although the similar SKU followed the same trend as the component group it was in, it did so at a much lower scale. An example of this is shown in a closer examination of Kaiser Component Group 6 and most similar SKU 27707 (Figure 11).

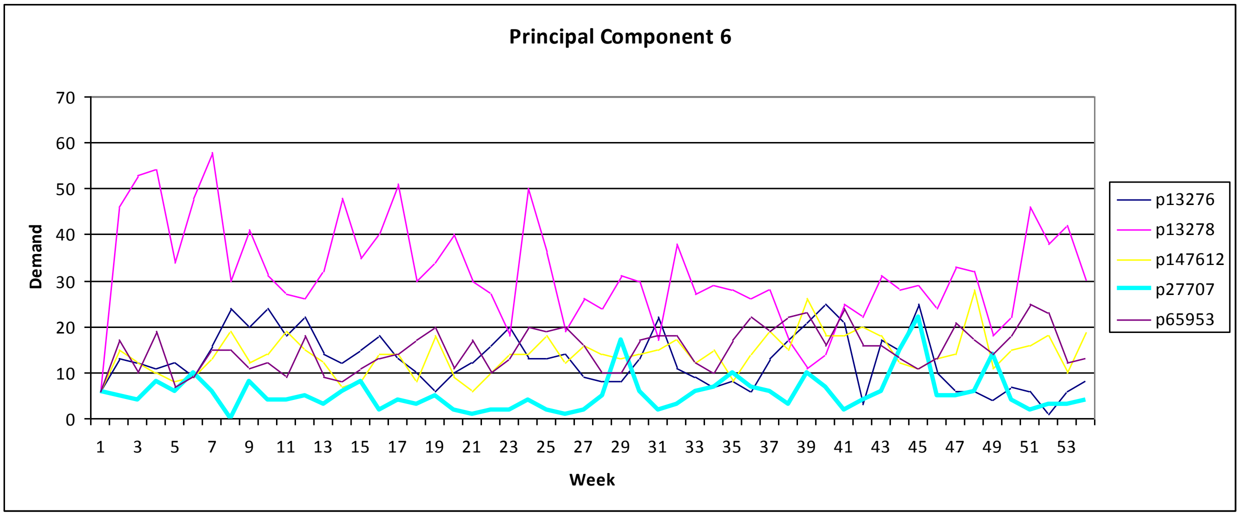

Figure 11: Component Group 6

Therefore, in an attempt to improve the PCA based forecasts, an additional step was performed in which the forecasts were scaled by a factor derived by dividing the component group's average forecast by the most similar SKU's average forecast (Table 31).

prodID	Top Like SKU	Top SKU avgDmd	KAI	KAI avg fcst	KAI scale	SCR	SCR avg fcst	SCR scale
p5452	5606	4	11	3	0.71	6	6	1.62
p5613	5601	2	5	4	1.59	7	3	1.27
p5615	5601	2	5	4	1.59	7	3	1.27
p13281	13313	4	2	5	1.23	6	6	1.44
p13348	13318	6	12	9	1.50	6	6	1.02
p13349	13313	4	2	5	1.23	6	6	1.44
p13350	13313	4	2	5	1.23	6	6	1.44
p13351	13313	4	2	5	1.23	6	6	1.44
p13393	13318	6	12	9	1.50	6	6	1.02
p13400	13318	6	12	9	1.50	6	6	1.02
p13401	13318	6	12	9	1.50	6	6	1.02
p13402	13318	6	12	9	1.50	6	6	1.02
p13403	13317	9	12	9	1.00	4	7	0.78
p27757	27707	6	6	16	2.85	5	9	1.64
p27758	27707	6	6	16	2.85	5	9	1.64
p27759	27707	6	6	16	2.85	5	9	1.64
p27762	27707	6	6	16	2.85	5	9	1.64

Table 31: Scaling Factors

New scaled forecast were then generated by dividing the scaling factor into the original Kaiser and Scree based forecasts (Table 32).

wk	SKU	Top Like SKU	KAI	KAI scale	SCR	SCR scale	KAI Scaled	SCR Scaled	Scaled Kai |Err|	Scaled Scree |Err|	Scaled Kai Err %	Scaled Scree Err %	Scaled Kai Acc	Scaled Scree Acc
6	p5452	5606	11	0.71	6	1.62	180	181	98	99	1.20	1.21	0.00%	0.00%
13	p5613	5601	5	1.59	7	1.27	82	94	166	154	0.67	0.62	33.00%	38.09%
12	p5615	5601	5	1.59	7	1.27	85	99	46	60	1.19	1.53	0.00%	0.00%
8	p13281	13313d	2	1.23	6	1.44	204	195	36	27	0.21	0.16	78.70%	83.98%
42	p13348	13318	12	1.50	6	1.02	57	87	40	70	2.38	4.09	0.00%	0.00%
40	p13349	13313	2	1.23	6	1.44	90	71	23	4	0.34	0.07	66.30%	93.46%
41	p13350	13313	2	1.23	6	1.44	84	67	14	31	0.14	0.31	85.89%	68.68%
41	p13351	13313	2	1.23	6	1.44	84	67	2	19	0.02	0.22	97.87%	78.27%
40	p13393	13318	12	1.50	6	1.02	73	101	18	46	0.33	0.84	67.26%	16.28%
40	p13400	13318	12	1.50	6	1.02	73	101	29	1	0.28	0.01	71.58%	99.07%
41	p13401	13318	12	1.50	6	1.02	65	95	10	40	0.18	0.73	82.14%	26.75%
50	p13402	13318	12	1.50	6	1.02	14	40	9	35	1.81	6.92	0.00%	0.00%
41	p13403	13317	12	1.00	4	0.78	97	74	54	31	1.27	0.73	0.00%	26.87%
22	p27757	27707	6	2.85	5	1.64	175	195	34	54	0.24	0.38	76.21%	62.03%
28	p27758	27707	6	2.85	5	1.64	143	161	8	26	0.06	0.20	94.17%	80.39%
22	p27759	27707	6	2.85	5	1.64	175	195	47	67	0.36	0.52	63.64%	48.02%
22	p27762	27707	6	2.85	5	1.64	175	195	17	37	0.10	0.23	89.53%	76.87%
						Σ	1855	2019	650	799	0.40	0.49	60.05%	50.87%

Table 32: Scaled PCA Based Class3id_4802 Error Report

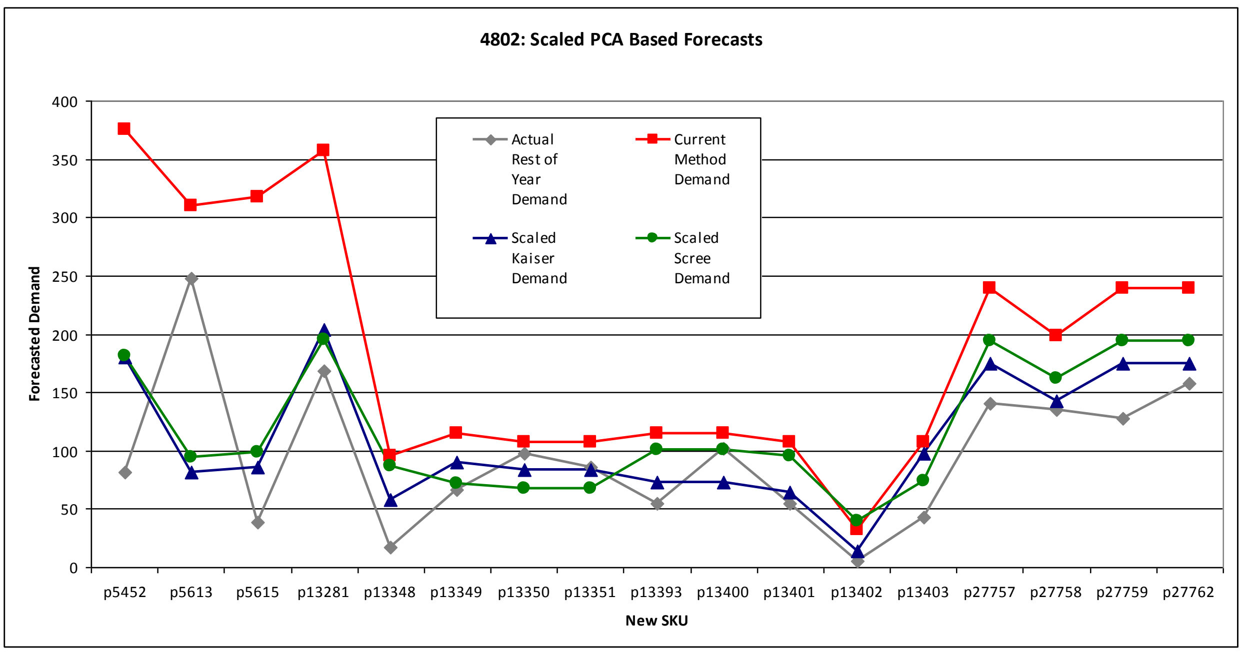

Figure 12: Forecasts Developed Using Scaled PCA Based Methods

Although the results look promising, prior to making any comparisons, the validity of making comparisons between the forecasting methods must be verified. That is, it must be stated with at least 95% confidence that the errors resulting from these forecasts belong to the same larger population. Two comparisons are made. The first is between the base error group and Kaiser based error group, which returned a p-value of 0.014. The second comparison was between the base error group and Scree based error group which returned a p-value of 0.034. Therefore, based on these results, it can be stated with 95% confidence that these two data sets are part of the same population and conclusions made from comparisons between the groups are valid

For Class3id 4802, the scaling factor significantly increased overall forecast accuracies, raising the Kaiser based forecasts to 60.05% and the Scree based forecasts to 50.87%. The overall scaling results, however, proved to be less promising. The Kaiser based forecast actually decreased in accuracies, across all five Class3ids, reducing to 13.28% forecast accuracy from 13.46%. The Scree based results faired slightly better with an increased overall forecast accuracy of 15.14%.

The above overall comparisons were validated with the hypothesis test comparing the current forecasting method error group against the PCA based forecast error groups. Resulting p-values of 0.023 and 0.014, for the scaled Kaiser and Scree based forecasts, respectively, validated the comparisons within the stipulated 95% confidence.

4.6 PCA and CJA: A Summary

	Acutal	Base (CJA)	Base (PCA)	Top SKU	Top 15% SKUs	Kaiser	Scree	Kaiser Scaled	Scree Scaled
Demand	21181	44930	41402	35187	39209	37636	35950	35349	34701
|Error|	-	24993	22038	18436	18871	18331	19073	18369	17975
Error	-	1.180	1.040	0.870	0.891	0.865	0.900	0.867	0.849
Accuracy	-	0	0	12.96%	10.91%	13.46%	9.95%	13.28%	15.14%

Table 33: Forecast Summary

Table 33 gives a summary of the forecast results for 129 new SKUs that appeared over 5 different Class3ids. The final results show that both CJA and PCA based forecasts improved current method forecasts, which produced 0% accuracies. Originally, the PCA based forecasts did not significantly improve upon the CJA based results, as was expected and in one case produced worse results. After a scaling factor was introduced, however, the forecasts improved slightly in accuracy, yet still not by great degrees (that is, only by a few percent).

5.0 Conclusion

The objective of this thesis was to improve upon currently employed methods in the forecasting of new SKUs. Current forecasting methods involve the processing of large amounts of historical demand data taken from seemingly similar existing SKUs (that is, Class3ids). It was found however that this current forecasting method resulted in very large errors with zero accuracy stemming from the great dissimilarities in demand trends that exist even within the same Class3id. Thus, to accomplish the objective, the goal was to refine the data set of existing SKUs from which the new SKU forecast would be drawn to encompass only data that would truly reflect the new SKU demand.

Two approaches were taken to reducing the data set. The first approach was to identify, quantitatively, existing SKUs that were similar to the new SKU according to consumer demand. The application of Conjoint Analysis, which assigned importance weights to SKU attributes based on consumer demand allowed for the calculation of SKU similarities and thus a means to identify the SKU which was the most “like” the new SKU. As expected, forecasts based on the CJA results produced forecasts with improved accuracies over current methods, with overall forecast accuracies of 12.96%.

An alternate approach to refining the original data set was to identify and isolate trends in the historical data and assimilate the new SKU forecast to one of the identified trends. Principal Component Analysis, a method used to identify relationships and patterns that exist in data of large dimensions, was utilized to identify time series demand trends across large quantities of SKUs. The PCA results were positive with demand trends clearly emerging from within the data set as a whole. For the next step, associating the new SKU with one of the emerging trends, the choice seemed obvious – the new SKU would assimilate the same demand trend as its most similar SKU, identified from previous CJA results. Loading the new SKU into its component group and performing forecasts returned poor results, with accuracies barely improving over CJA based forecasts in one case (at 13.46% accuracies) and not improving at all in the second instance (at 9.95%). A quick scaling factor was introduced and the PCA based forecasts showed some improvements with accuracies now up to 15.14% accuracies, a few percentage points above CJA based forecasts.

5.1 Further Research and Direction

The PCA based forecasts, although showing some improvements following the introduction of scaling factors, did not show significant improvements over the CJA based forecasts and thus did not perform as well as expected. For further study, it is believed that PCA based forecasts can be improved upon in two areas. First, alternative methods of loading the new SKU into a component group may result in the new SKU being loaded into a group that more closely resembles its demand trend. Second, further analysis of each component group may reveal improved scaling factors, the application of which will generate more accurate forecasts.

6.0 References

[1] "Demand Forecasting." Wikipedia. 28 October, 2008. <http://en.wikipedia.org/wiki/Demand_forecasting>.

[2] "The 8 Steps of the Forecasting Process Using Demand Planning Software." Applied Forecasting. 13 August 2008. 12 September 2008 <http://www.appliedforecasting.com/index.php?option=com_content&task=view&id=625>.

[3] Curry, Joseph. "Understanding Conjoint Analysis in 15 Minutes." Technical Papers. 1996. Sawtooth Software. 6 October 2008 <http://www.sawtoothsoftware.com/download/techpap/undca15.pdf>.

[4] Smith, Lindsay. "A Tutorial on Principal Component Analysis." COSC453. 26 February 2002. University of Otago Department of Computer Science. 6 October 2008 <http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf>

[5] Orme, B. "Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research." Madison: Research Publishers LLC, 2006.

[6] "Linear Regression and Multilinear Regression." Division of Biocomputing: HSC School of Medicine. 6 October 2008 <http://biocomp.health.unm.edu/biomed505/Course/Cheminformatics/advanced/data_classification_qsar/linear_multilinear_regression.pdf>

[7] "Principal Component and Factor Analysis." StatSoft, Inc. (2007). Electronic Statistics Textbook. Tulsa, OK: StatSoft. 6 October 2008 <http://www.statsoft.com/textbook/stathome.html>

[8] "Rotatefactors." Mathworks. 13 December 2008 < http://www.mathworks.com/access/helpdesk/help/toolbox/stats/index.html?/access/helpdesk/help/toolbox/stats/rotatefactors.html>