Inventory management using data mining: forecasting in retail industry

By

Peng Zhang

A Thesis Submitted to

Saint Mary's University, Halifax, Nova Scotia

in Partial Fulfillment of the Requirements for

the Degree of Master of Science in Applied Science (Computer Science).

December 15th, 2010, Halifax, Nova Scotia

Your file Votre référence
ISBN: 978-0-494-72013-4
Our file Notre référence
ISBN: 978-0-494-72013-4

Certification

Inventory management using data mining: forecasting in retail industry

by

Peng Zhang

A Thesis Submitted to Saint Mary's University, Halifax, Nova Scotia,
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Applied Science

December 17th, 2010, Halifax, Nova Scotia

Approved:	Dr. Pawan Lingras Supervisor Department of Mathematics and Computing Science
Approved:	Dr. Haiyi Zhang External Examiner Jodrey School of Computer Science Acadia University
Approved:	Dr. Norma Linney Supervisory Committee Member Department of Mathematics and Computing Science
Approved:	Dr. Hai Wang Supervisory Committee Member Department of Finance, Information Systems and Management Science
Approved:	Dr. Robert Dawson Program Representative
Approved:	Dr. Genlou Sun Graduate Studies Representative
Date:	December 17th, 2010

Acknowledgements

Sincere gratitude is given to all the people who made this thesis possible!

It is difficult to overstate my gratitude to my supervisor, Dr. Pawan Lingras. Without his knowledge, patience, inspiration and his great efforts including clear interpretation research findings, this thesis could not have been completed. He guided me through this interesting topic and introduced the world of research to me. His help during my thesis writing period has benefitted me in many ways, which includes clearly describing contents, summarizing contents from different perspectives and highlighting important points.

I gratefully acknowledge the examining committee members, Dr. Hai Wang, Dr. Norma Linney and the external examiner, Dr. Haiyi Zhang. Their involvements on the research helped me to think wider and further. Furthermore, I convey my appreciation to my friends, especially Professor Foster Lyne, who did proof readings and provided precious comments to my thesis. Their comments and support assisted me with this work through these remarkable and pleasant years.

I would like to thank the retail company for providing data for this study. I also want to highlight my thanks to Department of Mathematics and Computing Science, Faculty of Graduate Studies and Research, and Faculty of Science for providing me the opportunity to study at Saint Mary's University and the financial support. I really appreciate the help from people in the Department of Mathematics and Computing Science.

Finally, I would like to thank my parents, friends and my girlfriend Chuan Li for their unconditional support, patience and constant encouragements, which helped me to overcome every difficulty and accomplish this thesis.

Abstract

Inventory management using data mining: forecasting in retail industry

By Peng Zhang

Abstract: Inventory management, as an important business issue, plays a significant role in promoting business development. This study aims to apply data mining techniques, such as time series clustering and time series prediction techniques, in inventory management. Based on historical business data sets, time series clustering techniques, such as K-Means and Expectation Maximization are used to categorize inventories into reasonable groups. This study then identifies the most effective prediction technique to accurately predict inventory demands for each group. The traditional statistical evaluation metrics, such as Mean Absolute Percentage Error may not always be good indicators in an inventory management system, where the goal is to have as little inventory as possible without ever running out. The thesis proposes a more appropriated evaluation metric based on cost/benefit analysis of inventory forecasts. Results from a simulation program based on the proposed cost/benefit analysis are compared with statistical metrics.

Date: December 17th, 2010

Table of contents

Chapter 1 Introduction 1

1.1 Overview of retail industry 1

1.2 Overview of data mining and its applications 4

1.3 Objectives of the thesis 8

1.4 Organization of the thesis 10

Chapter 2 Literature review 11

2.1 Clustering 12

2.1.1 Static data clustering 12

2.1.2 Time series clustering 13

2.1.3 Clustering algorithms 14

2.1.3.1 K-Means 14

2.1.3.2 Hierarchical clustering 17

2.1.3.3 Self-organizing maps 19

2.1.3.4 Expectation Maximization (EM) 20

2.2 Inventory forecasting 21

2.2.1 Short-term forecasting 23

2.2.1.1 Regression 23

2.2.1.2 Exponential Smoothing 24

2.2.1.3 Autoregressive Integrated Moving Average (ARIMA) 26

2.2.1.4 Artificial neural networks (ANN) 27

2.2.2 Long-term forecasting 29

2.2.2.1 Recursive prediction strategy 29

2.2.2.2 Direct prediction strategy 30
2.3 Inventory forecasting evaluation 30
2.3.1 Statistical evaluation metrics 31
2.3.1.1 Mean square error (MSE) 32
2.3.1.2 Mean absolute deviation (MAD) 32
2.3.1.3 Mean absolute percentage error (MAPE) 32
2.3.2 Managerial evaluation metrics 32
Chapter 3 Study and experimental design 35
3.1 Overview of a small retail chain of specialty stores 35
3.1.1 Nature of the business 35
3.1.2 Available data 36
3.1.3 Overviews of business operations 36
3.1.4 Summary of sales distribution 53
3.2 Data preparation 53
3.2.1 Business understanding and data study 53
3.2.2 Data cleaning and extraction 57
3.2.3 Data consolidation 60
3.3 Data mining techniques used in inventory management 65
3.3.1 Products profiling and time series clustering 65
3.3.1.1 Stability analysis 65
3.3.1.2 Seasonality analysis 66
3.3.2 Inventory forecasting and time series prediction 68
3.3.2.1 Inventory forecasting models 68

3.4 Inventory forecasting evaluation ... 69

3.5 Simulation of business operations and performance ... 69

3.5.1 Inventory system operation simulation ... 70

3.5.2 Cost management ... 70

3.6 Experiment applications and softwares ... 71

Chapter 4 Product profiling and time series clustering ... 73

4.1 Stability analysis ... 73

4.1.1 A refined stability analysis ... 74

4.1.2 Time series clustering algorithms ... 77

4.1.3 Level-1 stability analysis – weekly analysis ... 78

4.1.3.1 Tier-1 weekly stability analysis ... 78

4.1.3.2 Tier-2 weekly stability analysis ... 80

4.1.4 Level-2 stability analysis – monthly analysis ... 82

4.1.4.1 Tier-1 monthly stability analysis ... 83

4.1.4.2 Tier-2 monthly stability analysis ... 85

4.1.5 Level-3 stability analysis – quarterly analysis ... 88

4.1.5.1 Tier-1 quarterly stability analysis ... 88

4.1.5.2 Tier-2 quarterly stability analysis ... 91

4.2 Seasonality analysis ... 94

4.2.1 Level-1 seasonality analysis – monthly analysis ... 95

4.2.1.1 Tier-1 monthly seasonality analysis ... 95

4.2.1.2 Tier-2 monthly seasonality analysis ... 102

4.2.2 Level-2 seasonality analysis –quarterly analysis ... 108

4.2.2.1 Tier-1 quarterly seasonality analysis 108

4.2.2.2 Tier-2 quarterly seasonality analysis 114

4.3 Summary of product analyses 120

Chapter 5 Inventory forecasting and business simulation 124

5.1 Inventory forecasting 124

5.1.1 Generic (global) optimal solutions 129

5.1.2 Local optimal solutions 131

5.1.3 Groups with strong sales patterns 136

5.1.4 Cross-year comparison 142

5.2 Business simulation 146

5.3 Summary of inventory forecasting and business simulation 152

5.3.1 Inventory forecasting and time series prediction 152

5.3.2 Business simulation 153

Chapter 6 Conclusions 155

6.1 Summary and Conclusions 155

6.2 Future directions 157

Appendix 160

References 176

List of Tables

Table 3-1: An overview of annual business operations 37

Table 3-2: An overview of quarterly business operations 38

Table 3-3: An annual overview of products 39

Table 3-4a: A quarterly overview of products (2005 and 2006) 40

Table 3-4b: A quarterly overview of products (2007) 41

Table 3-5: A monthly overview of products sold in 2005 42

Table 3-6: A monthly overview of products sold in 2006 43

Table 3-7: A monthly overview of products sold in 2007 44

Table 3-8: The quarterly distribution of products sold in 2005 45

Table 3-9: The quarterly distribution of products sold in 2006 45

Table 3-10: The monthly distribution of products sold in 2005 46

Table 3-11: The monthly distribution of products sold in 2006 47

Table 3-12: The distribution of single quarter selling products 48

Table 3-13: The distribution of single month selling products 49

Table 3-14: Invalid product samples in 2005 58

Table 3-15: Invalid product samples in 2006 59

Table 3-16: Sales records associated with product 068958048215 60

Table 3-17: Confusion of product IDs recognition in a CSV file 61

Table 3-18: Confusion of product IDs recognition in a CSV file 63

Table 3-19: Consolidated quarterly sales quantity data for time series clustering and time series prediction. 64

Table 4-1: A sample of time series clustering data sets 76

Table 4-2: Tier-1 weekly stability analysis with the EM algorithm in 2005 79

Table 4-3: Tier-1 weekly stability analysis with the EM algorithm in 2006 80

Table 4-4: Tier-2 weekly stability analysis with the EM algorithm in 2005 81

Table 4-5: Tier-2 weekly stability analysis with the EM algorithm in 2006 82

Table 4-6: Tier-1 monthly stability analysis with the EM algorithm in 2005 84

Table 4-7: Tier-1 monthly stability analysis with the EM algorithm in 2006 85

Table 4-8: Tier-2 monthly stability analysis with the EM algorithm in 2005 86

Table 4-9: Tier-2 monthly stability analysis with the K-Means algorithm in 2005 86

Table 4-10: Tier-2 monthly stability analysis with the EM algorithm in 2006 87

Table 4-11: Tier-2 monthly stability analysis with the K-Means algorithm in 2006 88

Table 4-12: Tier-1 quarterly stability analysis with the EM algorithm in 2005 89

Table 4-13: Tier-1 quarterly stability analysis with the EM algorithm in 2006 89

Table 4-14: Tier-1 quarterly stability analysis with the K-Means algorithm in 2005 90

Table 4-15: Tier-1 quarterly stability analysis with the K-Means algorithm in 2006 91

Table 4-16: Tier-2 quarterly stability analysis with the EM algorithm in 2005 92

Table 4-17: List of stable groups in 2005 and 2006 93

Table 4-18: Tier-1 monthly seasonality analysis in 2005 97

Table 4-19: Tier-1 monthly seasonality analysis in 2006 100

Table 4-20: Tier-2 2006 monthly seasonality analysis – Cluster 2 103

Table 4-21: 2005 seasonal groups based on monthly seasonality analysis..................... 106

Table 4-22: 2006 seasonal groups based on monthly seasonality analysis..................... 107

Table 4-23: Tier-1 2005 quarterly seasonality analysis 109

Table 4-24: Tier-1 2006 quarterly seasonality analysis 112

Table 4-25: Tier-2 2006 quarterly seasonality analysis - Cluster 2 115

Table 4-26: 2005 seasonal groups based on quarterly seasonality analysis 118

Table 4-27: 2006 seasonal groups based on quarterly seasonality analysis 119

Table 5-1: Multiple forecasting results based on product P101 in 2005 126

Table 5-2: An example of MAPE comparisons for group 05M10 128

Table 5-3: Generic solutions in 2005 and 2006 130

Table 5-4: The sample distribution of local optimal solutions 132

Table 5-5: MAPE comparison results for group 05Q20 134

Table 5-6: Comparison results of predicted quantity demands for P-4233149 in 05Q20 135

Table 5-7: MAPE comparison results for group 06M60 137

Table 5-8: Comparison results of predicted values for product P0114-3 138

Table 5-9: MAPE comparison results based on the month level prediction for group 05Q03 140

Table 5-10: MAPE comparison results based on the quarter level prediction for group 05Q03 141

Table 5-11: Cross-year comparison results for group 05MW 144

Table 5-12: Stable monthly and weekly products in 2005 and 2006.............................. 145

Table 5-13: MAPE comparison results based on product P114 148

Table 5-14: A sample report of business simulation based on product P114 149

Table 5-15: Sample distributions of local optimal solutions based on business simulation reports 151

Table A-1: The complete distribution of local optimal solutions based on MAPE 167

Table A-2: The complete distribution of local optimal solutions based on business simulation reports.. 175

List of Figures

Figure 1-1: Retail sectors year-over-year sales growth (2007-2009, by quarter) 2

Figure 1-2: Inventory turnover by retail trade group (2008) 3

Figure 1-3: Top drivers for retail distribution investment (2008) 4

Figure 1-4: A general data mining model in business industry 6

Figure 2-1: Clustering approaches: (a) raw-data-based, (b) feature-based and (c) model-based 14

Figure 2-2: K-means algorithms 17

Figure 2-3: Hierarchical algorithms 18

Figure 2-4: Topology of a simple self-organizing map 20

Figure 2-5: A suggested forecasting framework 22

Figure 2-6: A simple neural network model 28

Figure 3-1: The rank distribution of annual sales quantities in 2006 50

Figure 3-2: The rank distribution of annual sales profits in 2006 51

Figure 3-3: The frequency distribution of annual sales quantities in 2006 52

Figure 3-4: The frequency distribution of annual sales profits in 2006 52

Figure 3-5a: The database schema of the business data set 55

Figure 3-5b: The database schema of the business data set 56

Figure 3-6: Sample stability analysis result of Bread and Extreme 120C 66

Figure 4-1: Tier-1 monthly seasonality analysis - 2005 monthly sales trend 98

Figure 4-2: Tier-1 monthly seasonality analysis - 2006 monthly sales trend 101

Figure 4-3: Tier-2 monthly seasonality analysis – 2006 monthly sales trend of Cluster 2 104

Figure 4-4: Tier-1 seasonality analysis – 2005 quarterly sales trend 110

Figure 4-5: Tier-1 seasonality analysis – 2006 quarterly sales trend 113

Figure 4-6: Tier-2 seasonality analysis – 2006 quarterly sales trend of Cluster 2 116

Figure 4-7: Product distribution in 2005 122

Figure 4-8: Product distribution in 2006 122

Figure 4-9: Profits distribution in 2005 123

Figure 4-10: Profits distribution in 2006 123

Chapter 1

Introduction

The retail sector plays a key role in bridging production and consumption. Its productivity and success are strongly affected by some critical areas, such as, supply chain management (SCM), customer relationship management (CRM), key performance indicator (KPI) measurement, and e-commerce. Discovering strategic information and utilizing performance indicators are essential for retailers to make business decisions and improve business performance with intentions of developing a relatively high rate of return on investments.

1.1 Overview of retail industry

The retail sector is a vital part of Canada's economy and society. According to Industry Canada and Retail Council of Canada, the direct contribution of retail trade to the economy was $74.2B in 2009, representing 6.2% of Canada's gross domestic product (GDP) (Industry Canada and Retail Council of Canada, 2010; Aberdeen Group, 2010). The rate of Canada's retail sector GDP growth was 34% faster than the U.S. retail sector and 96% greater than the Canadian economy between 2004 and 2008. Retail employment grew 2.4% per year from 2002 to 2009 while employing 2.0 million people, or 11.9% of the total working population in 2009 (Industry Canada and Retail Council of Canada, 2010; United States Census Bureau, 2010). Figure1-1 shows Canadian and United State's retail sectors year-over-year sales growth from 2007 to 2009 by quarter.

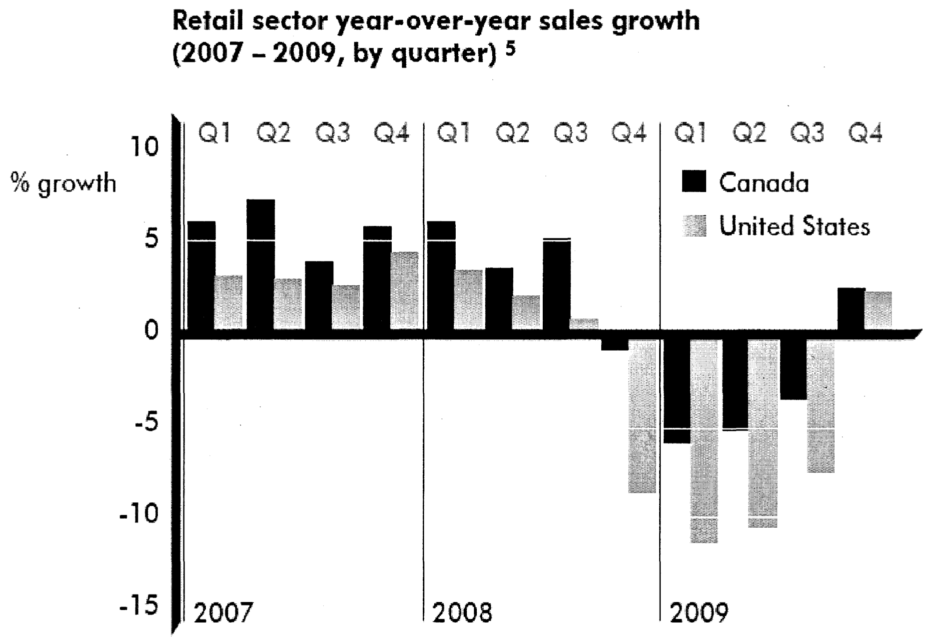


Figure 1-1: Retail sectors year-over-year sales growth (2007-2009, by quarter) (Industry Canada and Retail Council of Canada, 2010)

Inventory turnover[1], a primary KPI, measures how quickly the merchandise of a retailer is sold and replaced over a given time. That is, a higher turnover generally implies a lower holding cost for the retailer. Convenience and specialty food stores have the 2nd highest inventory turnover rate followed by Supermarkets. Figure 1-2 demonstrates the inventory turnover by retail trade group in 2008.

[1] Inventory turnover = COGS / Average inventory, where Average inventory = (Starting inventory + Closing inventory) / 2. (For example: 1 inventory turn is equal to a retailer having 365 days of inventory, 12 is 1 month of inventory, and 365 is 1 day of inventory.)

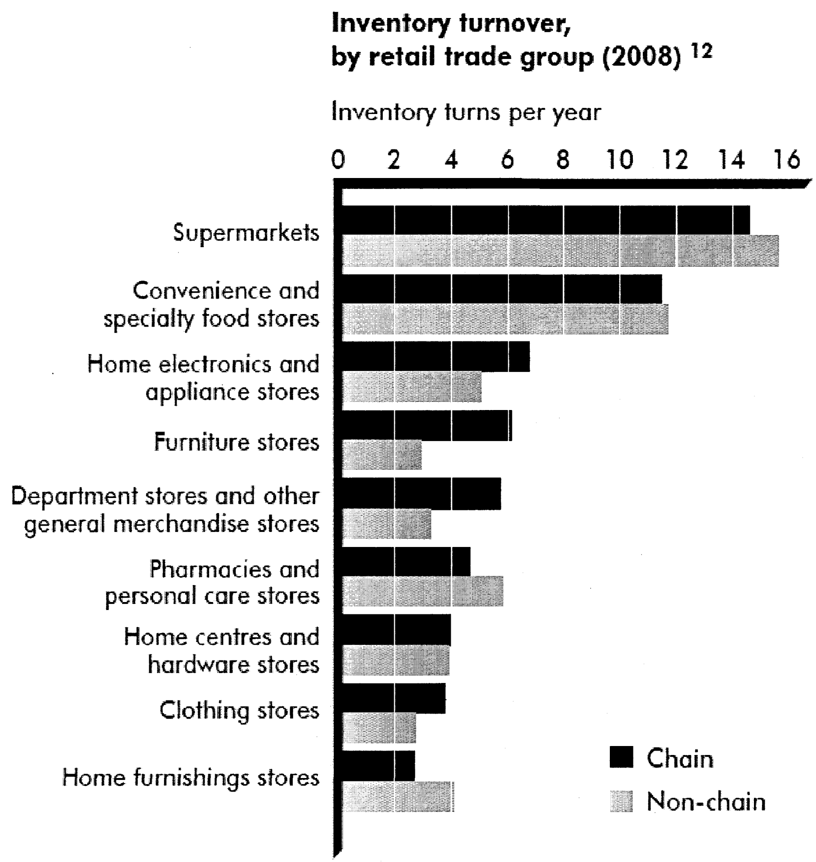


Figure 1-2: Inventory turnover by retail trade group (2008) (Industry Canada and Retail Council of Canada, 2010; Statistics Canada, 2010)

Cost of goods sold (COGS, i.e. procurement, landing, transportation, selling costs), non-revenue related expenses control and static merchandise pricing are the leading business pressures for North American retails firms. Figure 1-3 illustrates top drivers for retailer distribution investment.

Figure 1-3: Top drivers for retail distribution investment (2008) (Industry Canada and Retail Council of Canada, 2010)

1.2 Overview of data mining and its applications

Most, if not all, new information is digitalized in the modern world. Data, which carries information, is commonly collected and managed in many areas, for example, finance, economics, inventory management, weather forecast, military, scientific research and government. Due to the wide availability of huge amounts of data and imminent needs of turning data into information and knowledge, data analysis and information discovery from metadata attract significant attention of researchers.

Data mining, also known as Knowledge Discovery in Databases (KDD), is one of the fastest growing fields in computer science. It meets the rapidly increasing needs of information and knowledge discovery. Data mining works with multidisciplinary fields, for example, database technology, information retrieval, pattern recognition, machine learning, statistics, artificial intelligence, data visualization and high–performance computing (Han et al., 2006). In the business world, data mining is often used in areas of marketing analysis, fraud detection, risk analysis, and inventory management (Klosgen & Zytkow, 2002). Figure 1.4 demonstrates a general data mining model in business world. Information is discovered from data. It provides scenarios to support decision making. Finally, business decisions affect business performance. Feedbacks are also received from top to bottom.

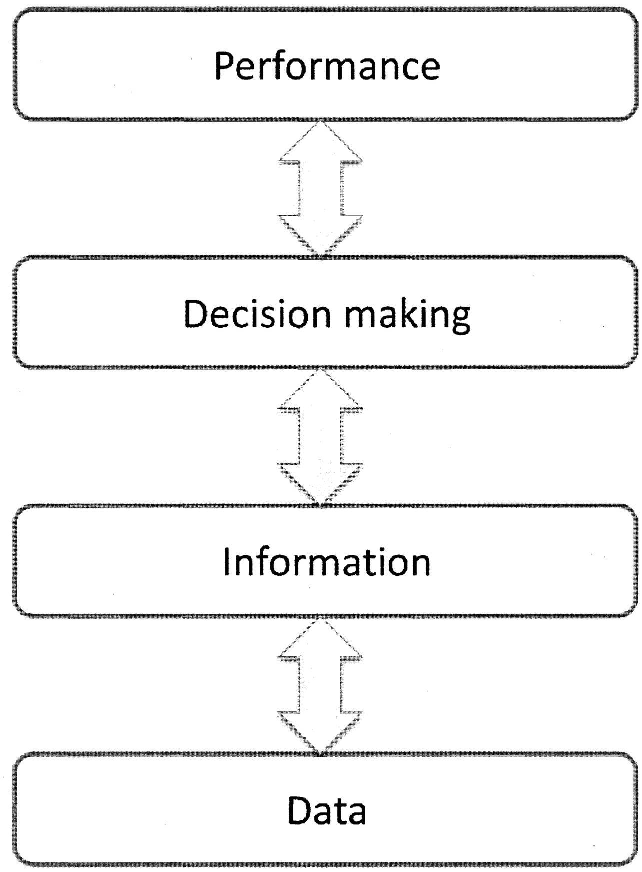


Figure 1-4: A general data mining model in business industry

Clustering, classification, association and prediction are some of the commonly used data mining techniques. Clustering techniques discover and identify data distributions and patterns from an unlabeled data set (Lingras & Akerkar, 2008). It is used to categorize data into homogeneous groups where similarity within a group is minimized and similarity between groups is maximized (Liao, 2005). Clustering approaches and methods can be applied in many ways. Based on the original data, two types of clustering approaches are distinguished. Static data clustering works on data that do not change with time or change negligibly (Liao, 2005). Time series clustering applies on data that change with time. Some popular time series clustering techniques will be described in section 2.1.3. Classification techniques classify data objects based on their values on certain attributes (Chen et al., 1996). Association techniques derive a set of strong association rules that discover local patterns in unsupervised learning systems (Kantardzic, 2003). Since association mining techniques may repeatedly scan through a large data set to find different association patterns, high-performance computing is an essential concern for huge amount of computations (Chen et al., 1996). Prediction techniques forecast future values based on historical data sets (Sorjamaa et al., 2007). Time series prediction techniques are often used on data that comprises of values that change over time (Lingras & Akerkar, 2008). Time series prediction models can be categorized into two groups: short-term prediction models and long-term prediction models. While short-term prediction models predict values of one period ahead, long-term prediction models predict values of multiple periods ahead (Sorjamaa et al., 2007). Some commonly used prediction techniques, such as, regression, exponential smoothing, autoregressive integrated moving average (ARIMA) and artificial neural networks, will be discussed in section 2.2.1.

Many commercial data mining applications have been developed to satisfy practical business requirements. These include SAS product - SAS Enterprise Miner (http://www.sas.com), Microsoft product – SQL Server Business Intelligence Development Studio (http://msdn.microsoft.com) and IBM products - SPSS (http://www.spss.com), DB2 Intelligent Miner (http://www.ibm.com/software/data/db2) and Cognos (http://www.ibm.com/software/data/cognos/). However, data mining applications are not self-sufficient. Strong analytical skills and comprehensive understanding of data are mandatory to perform data mining processes successfully.

1.3 Objectives of the thesis

Inventory is a significant portion of the current assets of any business enterprise (Kruger, 2005). Inefficient inventory management may cause a series of problems, for example, loss of productivity, inventory over-stocking and a reduction of customer commitment level. Any of these problems can be significant. On the other hand, efficient inventory management improves business performance and provides competitive advantages.

In modern business industry, inventory forecasting plays an essential role in inventory management system. Inventory forecasting predicts future demand of products. The goal of inventory management is to carry as little quantity as possible while satisfying all the sales requirements. Many researchers focus on finding a generic forecasting solution for all the products. However, products are distinguished by their seasonal sales patterns and volatilities in sales demand. One generic solution may not always be able to predict optimal quantity demand for each product.

In this study, we will use data mining techniques to improve inventory management. The first step is to use time series clustering techniques, using algorithms such as K-Means and EM, to categorize products into several reasonable groups based on product sales patterns. Secondly, we will apply some commonly used forecasting techniques on every product. The optimal forecasting solution for each product will be identified by comparison and evaluation of inventory forecasting models. Thus, inventory management will be improved simultaneously. Traditionally, inventory forecasting is evaluated by statistical indicators, such as, mean square error (MSE) and mean absolute percentage error (MAPE). However, statistical indicators may not always be reliable to identify the best fit forecasting solution. Particularly, they may not be able to provide inventory managers with reasonable criteria to make decisions. For example, a prediction error between actual value of 100 and predicted value of 101 is 1; the percentage error is 1%. On the other hand, a prediction error between actual value of 1 and predicted value of 2 is also 1; the percentage error is 100%. In this case, the percentage error is 99 times bigger than the one in previous situation. Although the difference is large in terms of percentage error, it has very little impact on the cost of inventory since the actually quantity difference is only 1.

Cost management is another critical factor that affects inventory management. Replenishment costs are relatively stable due to fixed administrative costs. It changes negligibly with the size of the order. Shortage costs and carrying costs can be volatile since they change with quantities of stock products. These costs are closely associated with inventory forecasting. Thus, appropriate managerial adjustments over inventory forecasting results are required to minimize inventory costs. Defining managerial adjustments for each product can be a challenge. In our study, a simulation program is proposed to support business decisions making. It simulates business operations based on inventory forecasting results and historical quantities of demand. Business reports regarding cost management, which will be generated from this program, provide managerial metrics to identify the optimal forecasting solution for each product. Such a simulation will make it possible to determine which inventory forecasting model and ordering strategy are appropriate for each product.

1.4 Organization of the thesis

This study focuses on applying data mining techniques to inventory management, especially inventory forecasting in wholesale and retail industry. Commonly used data mining methods and techniques in inventory forecasting and evaluation metrics are introduced in Chapter 2. These include time series clustering, prediction, and statistical as well as managerial evaluation metrics. The detailed goals of this study, description of the data sets, and experimental design are addressed in Chapter 3. Chapter 4 demonstrates a practical product profile analysis. It applies time series clustering techniques to products' stability and seasonality analyses. An empirical inventory forecasting experiment is demonstrated in Chapter 5. It applies time series prediction techniques. Time series forecasting evaluations are also conducted in Chapter 5. The demonstration in Chapter 5 confirms the usability of our comprehensive simulation program. Further discussion of using data mining techniques in inventory management is demonstrated in Chapter 6. Summary, conclusions and future research directions are also provided in this chapter.

Chapter2

Literature review

Inventory management (IM), as an essential business issue, plays a significant role in improving business performance. Efficient inventory management increases inventory accuracy, automates order process and optimizes business productivity. For over half a century, hundreds of books and journals were written about potential and actual uses of operation researches in inventory management (Silver, 1981). Inventory managers in most organizations are making decisions for large numbers of inventory items taking into consideration a diverse collection of factors (e.g., demand patterns, delivery methods and supply modes) and constraints in the areas of marketing, supplier, and internal capabilities (e.g., budget limitations, vendor restrictions, and desired customer service levels) (Silver, 1981; Hogarth & Makridakis, 1981). Time series prediction is an important aspect of effective inventory management. There are three key questions that affect making decisions on item-by-item basis in inventory management:

(i) "How often should the inventory status be determined, that is, what is the review interval?

(ii) When should a replenishment order be placed?

(iii) How large should the replenishment order be?"

In inventory management, a number of objectives are of interest to inventory managers. These include maximizing profits (with or without discounting), rates of return on investment and the chance of survival, minimizing cost (with or without discounting), ensuring flexibility of operations and determining feasible solutions. Silver (1981) described four categories of costs that relate to inventory management decision making, namely (i) replenishment costs, (ii) carrying costs, (iii) costs of insufficient supply in the short run, and (iv) system control costs (Silver, 1981; Peterson & Silver, 1979).

Many researchers have made progresses in these areas. In this chapter, we review their studies in three subsections: (1) clustering, (2) inventory forecasting, and (3) inventory forecasting evaluation.

2.1 Clustering

Clustering, perhaps the most frequently used data mining technique, is a convenient technique to discover data distribution and patterns in underlying data (Lingras & Akerkar, 2008). The goal of clustering is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the object similarity within groups is minimized and the object dissimilarity between groups is maximized (Liao, 2005).

2.1.1 Static data clustering

Data sets are called static if all their feature values do not change with time, or change negligibly (Liao, 2005). Extraordinary amounts of clustering analysis have been performed on static data sets. These are called static data clustering analysis. Most clustering programs, which were developed as an independent program or as part of a large suite of data analysis or data mining software, work primarily with static data set. Han, Kamber and Pei (2006) classified various static data clustering methods into five major categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods (Han et al., 2006).

2.1.2 Time series clustering

In contrast to static data, time series data is comprised of values that change with time (Liao, 2005). Time series data is pervasive in various areas, such as, science, engineering, business, finance, economics, health care and government. Compared to studies on static data, the number of time series data researches is relatively scant. However, researchers are paying increased attention to time series data. Works in this area can be classified into two main categories: whole sequence clustering and sub-sequence clustering (Keogh & Lin, 2005). Whole sequence clustering is the time series clustering analysis performed on a set of individual time series data. In sub-sequence clustering, given a single time series, sub-sequences are extracted from a sliding window. Sub-sequence clustering is then performed on the extracted time series.

Given a set of unlabeled time series data, the choice of clustering algorithm depends on the type of data available, the purpose of clustering and particular applications (Liao, 2005). As far as time series data is concerned, distinctions can be made based on the nature of data, such as, whether the time series are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or multivariate, and equal or unequal in length. Figure 2-1 outlines the three different approaches: raw-data-based, feature-based, and model-based. The raw-data-based approach works directly on raw data. The feature-based and model-based approaches transform the raw data into feature based vectors and model based parameters before clustering algorithms are applied (Silver et al., 1998).

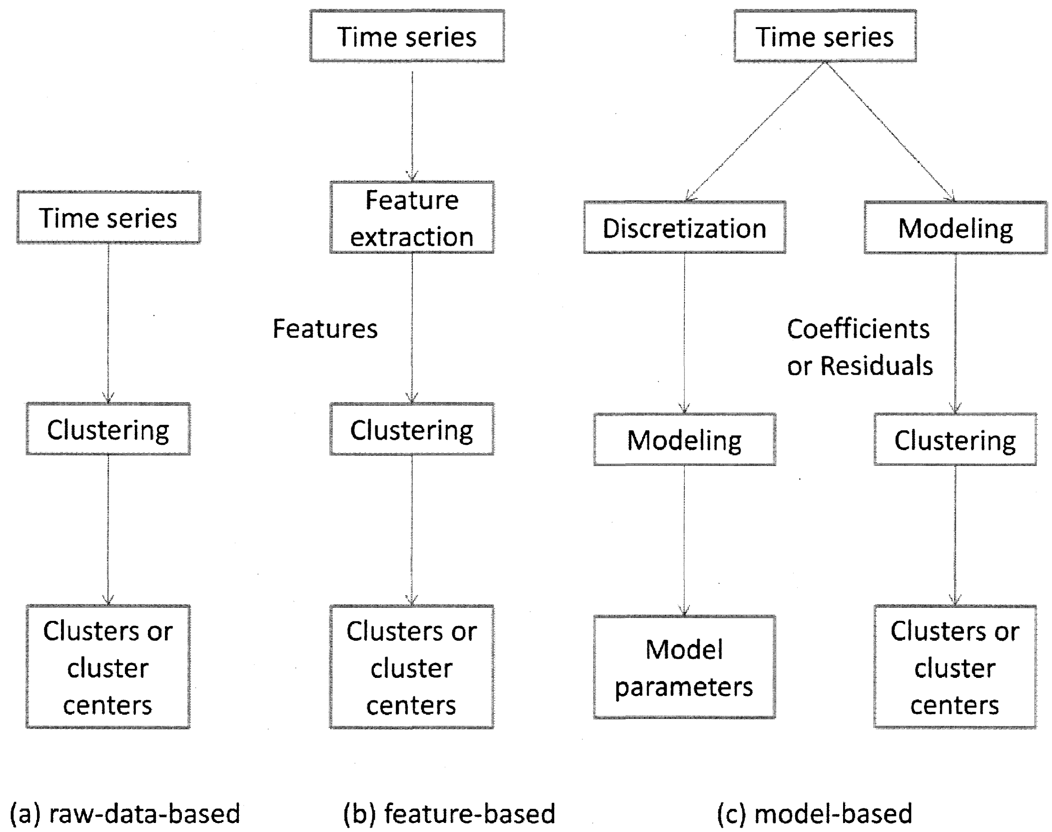


Figure 2-1: Clustering approaches: (a) raw-data-based, (b) feature-based and (c) model-based (Liao, 2005)

2.1.3 Clustering algorithms

There are many clustering methods, specifically partitioning methods, hierarchical methods, and model-based methods that are directly utilized or modified for time series clustering. Several popular algorithms and procedures are reviewed in this Section.

2.1.3.1 K-Means

K-means algorithm, published by MacQueen (1967), is the most commonly used clustering algorithm in practice (Lingras & Akerkar, 2008; Larose, 2005). This algorithm has an input of predefined number of clusters, which is called *k*. “Means” stands for the average location of all the members of a single cluster. The main goal of it is to minimize the objective function, which is normally defined as the total distance between all patterns from their respective cluster centers (Liao, 2005). The algorithm proceeds as follows (Larose, 2005).

Step 1: The expected number of clusters is required as an input from the user. This number is usually denoted as *k*.

Step 2: Randomly assign *k* data objects as the initial cluster centers.

Step 3: For each data object, the nearest cluster center is to be discovered. Usually, the distance, denoted by *d*, is used to identify the nearest cluster center. It is calculated as the Equation 2-1.

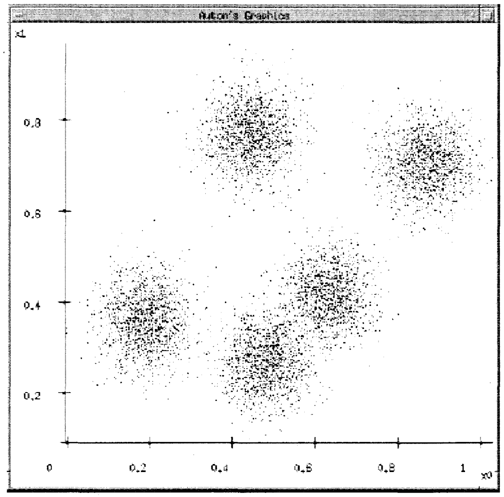
$$d = \| x_i - c_j \| \qquad \text{Equation 2-1}$$

where x_i is the data object and c_j is the cluster center, which corresponds to the minimal values of d. It represents the cluster center that x_i belongs to. Hence, the j is the cluster number that the data object belongs to and $1 \le j \le k$. Thus, in a sense, each cluster center has a subset of records, thereby representing a partition of the data sets. Therefore, we have the data objects divided into *k* clusters, C_1, C_2, …, C_k.

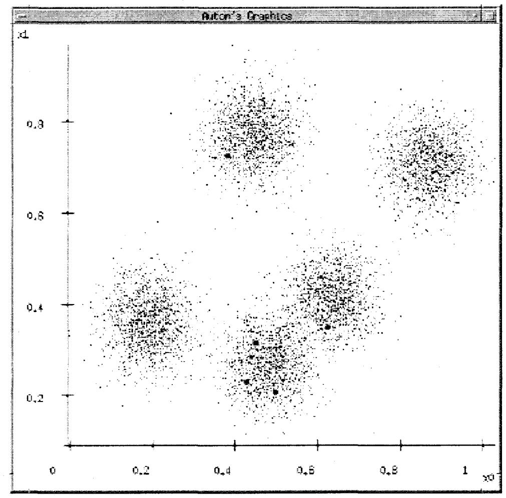
Step 4: For each of the *k* clusters, the new cluster center is to be located. Suppose that we have *n* data objects (a_1, b_1, c_1), (a_2, b_2, c_2), …, (a_n, b_n, c_n), the new cluster center is the mean of these data objects, which is $(\sum a_i/n, \sum b_i/n, \sum c_i/n)$.

Step 5: Repeat step 3 to step 5 until the location of cluster centers have no more changes.

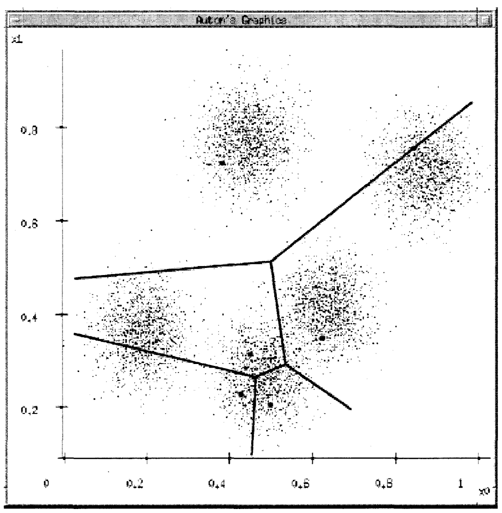
Figure 2-2 demonstrates the process of K-means algorithm.



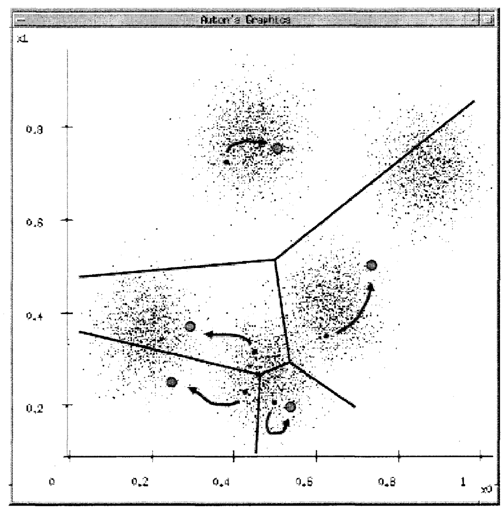
1. Original data set



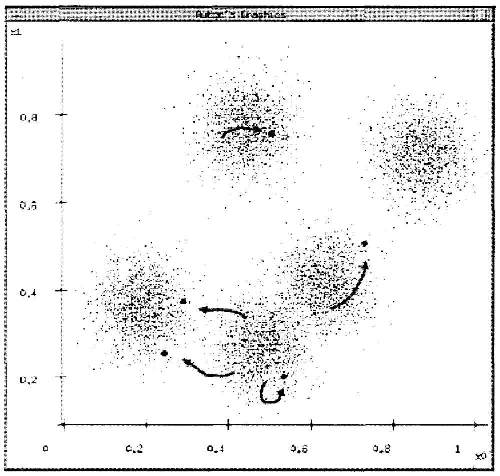
2. Randomly pick 5 data points



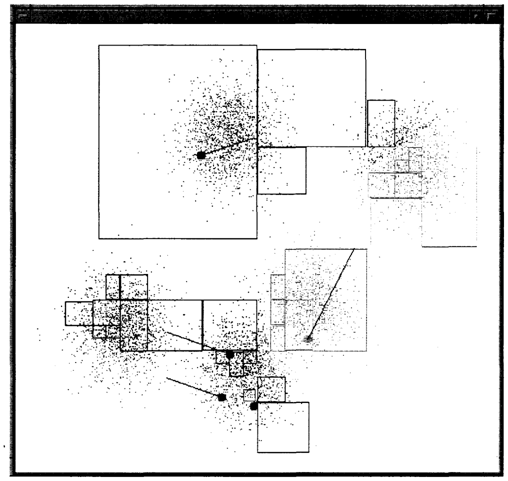
3. Divide data into 5 clusters by distances



4. Find new center of 5 clusters



5. Start clustering as step 2



6. Repeat steps 3 to 5 until no changes

Figure 2-2: K-means algorithms

2.1.3.2 Hierarchical clustering

Hierarchical clustering algorithm, proposed by Joe H. Ward in 1963, is a widely used algorithm that forms hierarchical groups of mutually exclusive subsets (Eisen, 1998; Jain & Dubes, 1998; Ward, 1963). Hierarchical clustering algorithms group data objects (time series here) into a tree of clusters (Liao, 2005). Two types of hierarchical clustering methods are often used: agglomerative and divisive. They are distinguished by their clustering strategies. The agglomerative hierarchical clustering algorithm is performed with a bottom-up strategy, while the divisive hierarchical clustering algorithm follows a top-down strategy. The agglomerative hierarchical clustering algorithm is more popular than the divisive algorithm. The algorithm proceeds as follows (Liao, 2005; Ward, 1963).

Step 1: Each object form a cluster that has only one object. Thus, *n* objects start with *n* clusters.

Step 2: Estimate the distance between two clusters as the distance of the closest pair of data points belonging to different clusters, and then merge the two clusters that have the minimum distance. The distance, denoted by $D(X, Y)$, between clusters X and Y can be estimated by the linkage function:

$$D(X, Y) = \min(d(x,y)) \quad \text{Equation 2-2}$$

where $d(x,y)$ is the distance between elements $x \in X$ and $y \in Y$.

Step 3: Repeat step 2 until all the objects are in a single cluster or until certain termination conditions are satisfied.

Figure 2-3 graphically demonstrates a simple model of hierarchical algorithm.

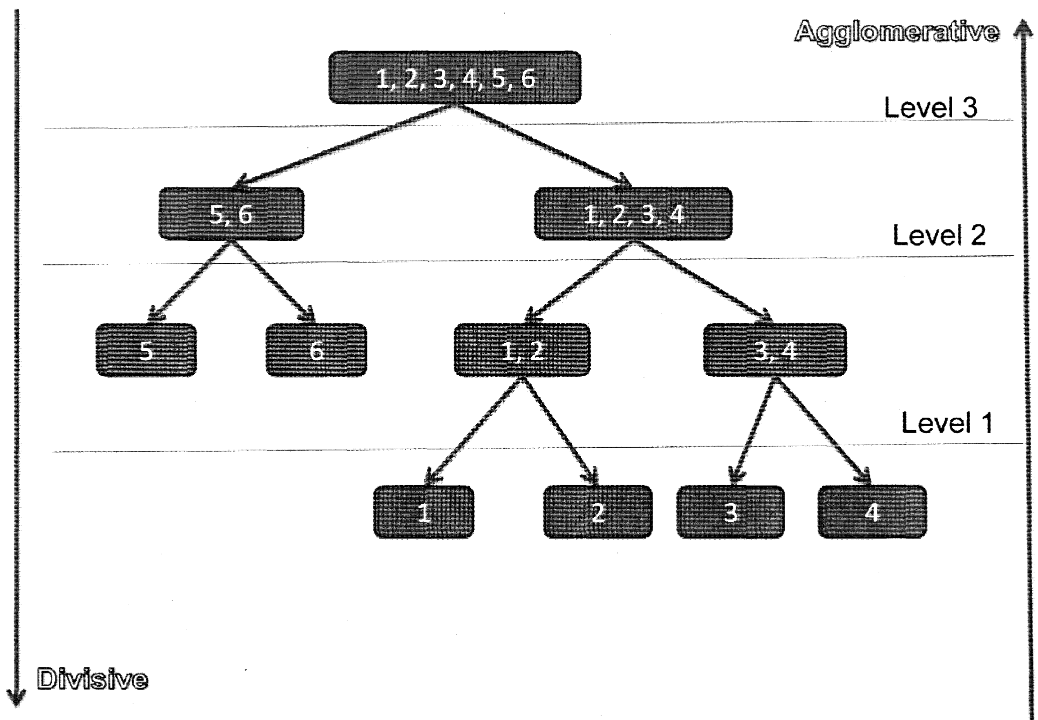


Figure 2-3: Hierarchical algorithms

2.1.3.3 Self-organizing maps

In 1979, Teuvo Kohonen first introduced Kohonen networks. It is also named as self-organizing maps (SOM) based on Neural Network models (Tamayo, 1999; Kohonen, 1979). It is one of the most commonly used unsupervised neural network models (Lingras & Akerkar, 2008). The goal of SOM is to convert a complex high-dimensional input signal into a simpler low-dimensional discrete map (Haykin, 1994). A SOM uses competitive learning steps and consists of a layer of input units, each of which is fully connected to a set of output units, which are arranged in some topology (the most common choice is a two-dimensional grid) (Lingras & Akerkar, 2008). The algorithm proceeds as follows (Lingras & Akerkar, 2008; Larose, 2005).

Step 1: For each output node j, the value $D(w_j, x_n)$ of the scoring function is calculated by Equation 2-3, where x_n is an input vector, w_i is the weight vector to node j.

$$D(w_j, x_n) = || x_n - w_j || \qquad \text{Equation 2-3}$$

Find the winning node J that minimizes $D(w_j, x_n)$ among all output nodes.

Step 2: Adjust the weights as:

$$w_{ij,\ new} = w_{ij,current} + h_{ck}(j) \times (x_{ni} - w_{ij,current}) \qquad \text{Equation 2-4}$$

The h_{ck} represents the neighbourhood function associated with the learning rate and x_{ni} that signifies the nth input to node J.

Step 3: Repeat step1 to step 2 for all the objects. Adjust the neighbourhood function.

Step 4: Repeat step 1 to step 3 for a specified number of epochs.

Figure 2-4: demonstrates the topology of a simple self-organizing map.

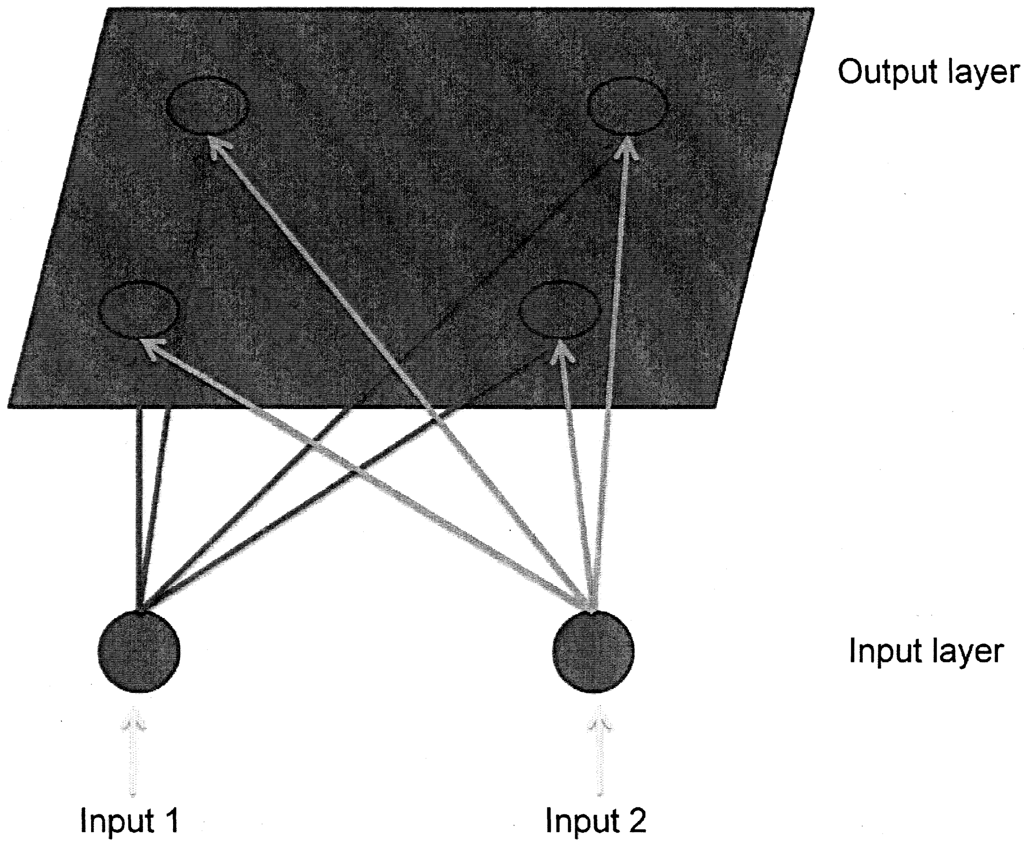


Figure 2-4: Topology of a simple self-organizing map (Larose, 2005)

2.1.3.4 Expectation Maximization (EM)

The expectation maximization algorithm, explained and named in 1977, is an effective and popular technique for estimating the parameters in statistical models (Bradley et al., 1998; Dempster et al., 1977). The EM algorithm iteratively refines an initial cluster model to better fit the data and terminates at a solution which is locally optimal or satisfies the underlying clustering criterion (Dempster et al., 1977; Fayyad et al., 1996; Bishop, 1995). The algorithm is computationally intensive and proceeds as follows (Bradley et al., 1998; Dempster et al., 1977).

Given a data set that contains a set of observed data, denoted by M, a set of unobserved latent data N, and a vector of unknown parameters θ, along with a likelihood function

$$L^{\theta}= L(\theta;M,N)=p(M,N|\theta) \qquad \text{Equation 2-5}$$

the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data:

$$L(\theta;M)=p(M|\theta)=\textstyle\sum_N L(M,N|\theta) \qquad \text{Equation 2-6}$$

The EM algorithm applies two steps to iteratively find the MLE.

Expectation step (E-step): find the expected value of the log likelihood function:

$$Q(\theta|\theta_t)=\varepsilon_{N|M,\theta t}[\log L(\theta;M,N)] \qquad \text{Equation 2-7}$$

Where Q is the quantity of the vector of parameters and ε is a stopping tolerance.

Maximization step (M-step): find the vector of parameters that maximizes the quantity in E-step:

$$\theta_{t+1}=arg_{\theta}^{maxQ(\theta|\theta t)} \qquad \text{Equation 2-8}$$

Stopping criteria: the calculation stops when $|L^{\theta}{}_{t}-L^{\theta}{}_{t+1}|<=\varepsilon$.

2.2 Inventory forecasting

Inventory forecasting predicts the future inventory demand based on historical and current demand (Sorjamaa et al., 2007). The demand quantity of a particular item or a group of related items can be considered as a time series of separate values (Silver et al., 1998). For example, the daily demand of milk for a year is a set of time series data that has 365 data objects. Inventory forecasting leads a critical path to support decision making in inventory management. Ideally, effective forecasts are performed with a combination of statistical forecasting and informed judgements, such as, promotions, competitor reactions and general economic factors. Figure 2-5 demonstrates a suggested forecasting framework. In this section, many popularly used time series forecasting models are reviewed.

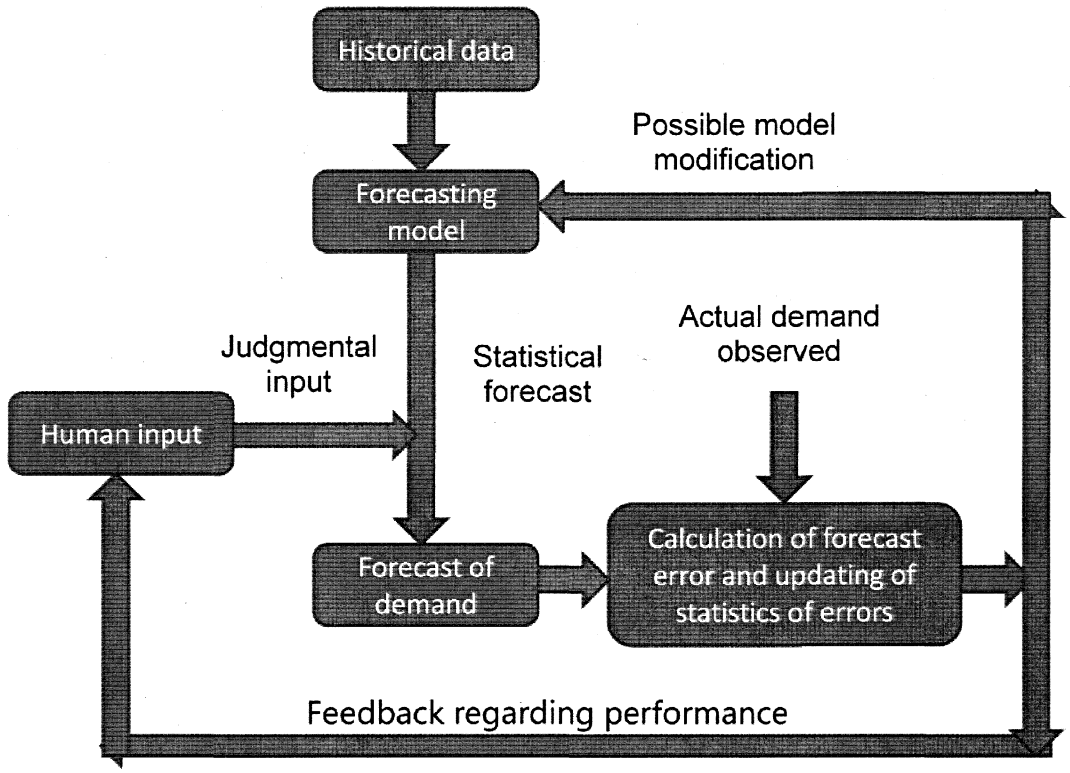


Figure 2-5: A suggested forecasting framework (Silver et al., 1998)

Conceptually, there are three steps involved in time series forecasting (Silver et al., 1998).

Step 1: Select an appropriate underlying model of the value pattern through time.

Step 2: Select the values for the parameters inherent in the model.

Step 3: Use the model (selected in Step 1) and the parameter values (chosen in Step 2) to forecast the future values.

Sorjamaa et al. (2007) claimed that there are two main types of time series forecasting: short-term and long-term forecasting (Sorjamaa et al., 2007). While short-term forecasting refers to one period ahead prediction, long-term forecasting refers to multiple periods ahead predictions. We will review some commonly used models of short-term forecasting in section 2.2.1 and strategies of long-term forecasting in section 2.2.2.

2.2.1 Short-term forecasting

Many time series prediction techniques are available to forecast future values of time series data (Sorjamaa et al., 2007). Based on these techniques, some short-term forecasting models are built to predict one period ahead value. In this section, we review some commonly used models for short-term time series forecasting.

2.2.1.1 Regression

Regression analysis, first introduced by Francis Galton (1886) and extended by G. U. Yule et.al (1897 and 1903), consists of graphic and analytic methods to discover relationships between one variable, named as a response variable or dependent variable, and one or more other variables, called predictor variables or independent variables (Lingras et al., 2008; Galton, 1886; Yule, 1897; Pearson et al., 1903). Regression analysis is one of the most widely used statistical tools because it establishes simple functional relationships among variables (Chatterjee & Hadi, 2006).

Linear regression with one input variable is the simplest form of regression. It models a response variable Y as a linear function of a predictor variable X. Given n samples or data points of the form (x_1, y_1), (x_2, y_2), …, (x_n, y_n), where $x_i \in X$ and $y_i \in Y$, a linear regression can be expressed as Equation 2-9:

$$Y = a + bX \quad \text{Equation 2-9}$$

where a and b are regression coefficients. Multiple regression is an extension of linear regression. It can be fed with two or more predictor variables. The response variable Y is modeled as a linear function of several predictor variables. In general, Equation 2-9 is extended to Equation 2-10:

$$Y = a + \sum_{i=1}^{n} b_i X_i \quad \text{Equation 2-10}$$

where n stands for a number of predictor variables, a and b_i are regression coefficients, and X_i is a predictor variable.

Once such an expression is obtained, the relationship can be utilized to predict values of the response variable, identify which variables mostly affect the response, or verify hypothesized causal models of the response (Mendenhall & Sincich, 1995). The regression finds a linear combination of input variables so that the sum of square errors is minimized between observed and predicted values.

2.2.1.2 Exponential Smoothing

Exponential smoothing is a widely used technique in time series forecasting (Geurts & Ibrahim, 1975; Lim & McAleer, 2001). Depending on the features of time series such as trend and seasonal effects, exponential smoothing models are grouped into many categories: simple exponential smoothing, exponential smoothing for linear trend (Holt's 1957 and Brown's 1959), and seasonal exponential smoothing (Silver et al., 1998; Billah et al., 2006; Taylor, 2003; Harrison, 1967; Gardner et al., 1989). These exponential smoothing models naturally relate to the error correction versions of exponential smoothing and underpin weighted average versions of the method (Gardner, 1985).

- Local Level Model (LLM):

$$y_t = L_{t-1} + \varepsilon_t \qquad \text{Equation 2-11}$$

  where L_t is a local level governed by the recurrence relationship $L_t = L_{t-1} + a\varepsilon_t$, where $0 \leq a \leq 1$. It underpins the simple exponential smoothing method (Brown, 1959).

- Local Trend Model (LTM):

$$y_t = L_{t-1} + b_{t-1} + \varepsilon_t \qquad \text{Equation 2-12}$$

  where b_t is a local growth rate. The local level and local growth rates are governed by the Equations $L_t = L_{t-1} + b_{t-1} + a\varepsilon_t$ and $b_t = b_{t-1} + \varepsilon_t$, respectively, where $0 \leq a \leq 1$and $0 \leq b \leq 1$. Note that $a' = [a\ b]$. This model underpins trend corrected exponential smoothing (Holt, 1957).

- Additive Seasonal Model (ASM):

$$y_t = L_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t \qquad \text{Equation 2-13}$$

  where s_t is the local seasonal component. The local level, growth and seasonal components are governed by $L_t = L_{t-1} + b_{t-1} + a\varepsilon_t$, $b_t = b_{t-1} + \varepsilon_t$, and $s_t = s_{t-m} + \varepsilon_t$, respectively, where m is the number of seasons in a year, $0 \leq a \leq 1$, $0 \leq b \leq 1$, and $0 \leq c \leq 1-a$. In this case,

  $s_t' = [L_t\ b_t\ c_t\ .\ .\ .\ c_{t-m+1}]$ and $a' = [a\ b\ c\ 0.\ .\ .0]$. This model is the basis of Holt–Winters' additive method (Winters, 1960).

In 2000, Chen et al. applied simple exponential smoothing model to quantify the bullwhip effect for supply chain systems (Chen et al., 2000). Experimental results clearly indicated that the bullwhip effect can be reduced by the retailer's sales forecast. In 2002, Snyder et al. conducted the study of exponential smoothing models with single source of error and multiple source of error on weekly sales figures. This showed that exponential smoothing remains appropriate under general conditions, where the variance is allowed to grow or contract with corresponding movements in the underlying level (Snyder et al., 2002).

2.2.1.3 Autoregressive Integrated Moving Average (ARIMA)

Autoregressive integrated moving average model, a variant of regression analysis model, is generally referred to as an ARIMA (p,d,q) model where p, d, and q (integers greater than or equal to zero) refer to the autoregressive, integrated, and moving average parts of the model respectively (Johnson & Thompson, 1975; Ray, 1982; Aviv, 2003; Contreras et al., 2003; Van Der Voort et al., 1996). Given a time series of data X_t, where t is an integer index, the X_t are real numbers, and L is the lead time. ARIMA models are used for observable non-stationary processes X_t that have some clearly identifiable trends:

- constant trend (i.e. a non-zero average) leads to $d = 1$
- linear trend (i.e. a linear growth behaviour) leads to $d = 2$
- quadratic trend (i.e. a quadratic growth behaviour) leads to $d = 3$

The non-stationary expression is:

$$Y_t = (1 - L)^d X_t \qquad \text{Equation 2-14}$$

The wide-sense stationary expression is:

$$\left(1 - \sum_{i=1}^{p} \gamma_i L^i\right) Y_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t \qquad \text{Equation 2-15}$$

where γ_i and θ_i are parameters. Standard forecasts model can be formulated for the process Y_t. Thus, X_t can be forecasted by integrating steps above.

In 2003, Contreras et al. successfully applied ARIMA models on time series data of hourly electricity price to forecast the future value of electricity price (Johnson & Thompson, 1975). In 1982, Ray modeled monthly sales forecasting of chemical food in an inventory control system with ARIMA (Ray, 1982).

2.2.1.4 Artificial neural networks (ANN)

Neural networks have been widely used to predict future values in investments, medicine, science, engineering, marketing, manufacturing and management (Lawrence, 1993). A neural network is a parallel distributed information processing system that consists of processing elements interconnected together with unidirectional signal channels called connections (Hecht-Nielsen, 1988). Each processing element has a single output connection, which branches into as many collateral connections as desired. Each collateral connection carries the same signal that is output by the processing element. The processing elements can output signals in any desired mathematical type. All of the processing that goes on within each processing element must be completely local, that is, it must depend only upon the current values of the input signals at the processing element through impinging connections and values stored in the processing element's local memory. Two important issues need to be addressed: the frequency at which data should be sampled and the number of the data points to be sampled. Figure 2-6 gives a simple structure of neural networks, where x $(t-2)$, x $(t-1)$ and x (t) are the input values at periods $(t-2)$, $(t-1)$, and t, x $(t+1)$ is neural network output. The standard neural network method of performing time series prediction is to induce the function f using any feed forward function approximating neural network architecture, such as, a standard MLP, an RBF architecture, or a Cascade correlation model (Gershenfeld & Weigend, 1993; Frank et al., 2001).

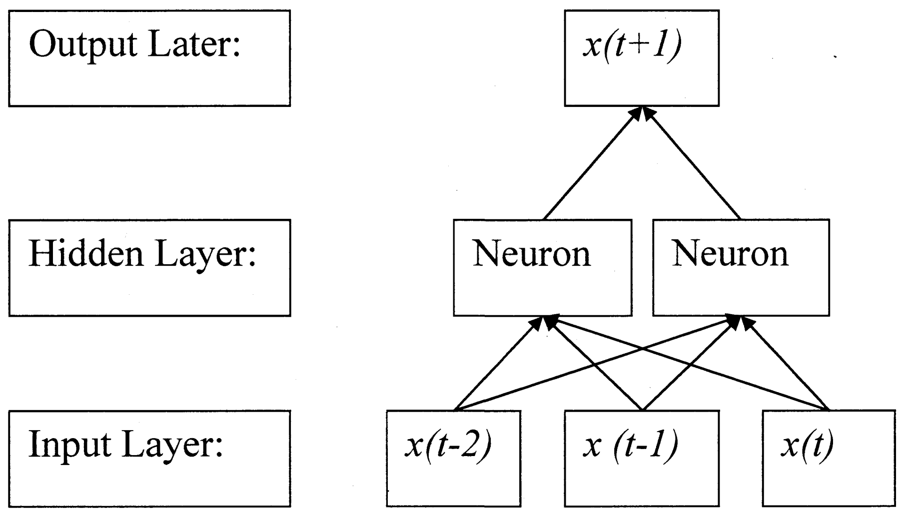


Figure 2-6: A simple neural network model

In 1994, recurrent neural networks model, considered to be a special case of nonlinear autoregressive moving average models, was successfully applied on data sets of daily and hourly electric loads from November 11, 1990 to March 31, 1991 to predict electric loads (Connor et al., 1994). In 2008, Lingras et al. applied time delay neural networks model to predict hourly traffic volume based on historical traffic data sets from two highway agencies in North America. Reasonable prediction results are obtained (Lingras et al., 2008). In 2002, Tseng et al. proposed a hybrid forecasting model - SARIMABP, which combines the seasonal time series ARIMA (SARIMA) and the neural network back propagation (BP) models (Tseng et al., 2002). Two seasonal time series data sets were experimented with: monthly sales value of the soft drink industry and monthly production value of Taiwan machinery industry from 1991 to 1996. The proposed model outperformed SARIMA and neural network models.

2.2.2 Long-term forecasting

Many strategies are available to build long-term forecasting models, which predict multiple periods ahead values (Sorjamaa et al., 2007). In the following sections, two variants of long-term forecasting strategies are reviewed: the direct and the recursive prediction strategies.

2.2.2.1 Recursive prediction strategy

Recursive prediction strategy seems to be the most intuitive and simple method to build long-term prediction models (Sorjamaa et al., 2007). The predicted values are used as predictor variables to predict the next ones. The detailed process is shown below:

Step 1: predict one period ahead value with a selected forecasting technique at period t, where X_{t-m}, $X_{t-m+1},\ldots, X_{t-1}$ are the predictor variables and m is the number of predictor variables used in the model,

$$X_t=f(X_{t-m}, X_{t-m+1},\ldots, X_{t-1}) \quad \text{Equation 2-16}$$

Step 2: X_{t-m+1}, $X_{t-m+2},\ldots, X_{t-1}$ and X_t, which is the predicted value from step 1, are used as predictor variables to predict value at period $t+1$,

$$X_{t+1}=f(X_{t-m+1}, X_{t-m+2},\ldots, X_{t-1}, X_t) \quad \text{Equation 2-17}$$

Step 3: $X_{t-m+2},\ldots, X_{t-1}, X_t$ and $X_{t+1},$ which is the predicted value from step 2, are used as predictor variables to predict value at period $t+1$,

$$X_{t+2}=f(X_{t-m+2},\ldots, X_{t-1}, X_t, X_{t+1}) \qquad \text{Equation 2-18}$$

Step 4: repeat forecasting process as previous steps until X_{t+N}, which is N periods ahead prediction value, is predicted.

In general, it is simple to process predictions with recurrent strategy. However, the calculation carries the errors in predicted values to next step. Therefore, using the predicted values as inputs variables deteriorates the accuracy of the prediction.

2.2.2.2 Direct prediction strategy

Another commonly used strategy to build long-term forecast models is the direct strategy (Sorjamaa et al., 2007). For the N periods ahead prediction, the model is

$$\widehat{X}_{t+n}=f_n(X_{t-m}, X_{t-m+1},\ldots, X_{t-1}) \quad \text{with } 1\leq n\leq N \qquad \text{Equation 2-19}$$

In this way, only original predictor variables X_{t-m}, X_{t-m+1},... and X_{t-1} are used to predict values at period $t+N$. The errors are not accumulated from period by period. However, N different models must be built if all the values from $\widehat{X}_{t+1}$ to $\widehat{X}_{t+N}$ need to be predicted. Compared to recursive prediction strategy, the direct strategy increases the complexity of prediction and improves the accuracy of forecast results.

2.3 Inventory forecasting evaluation

Production and operations managers repeatedly express that forecasting is a critical activity since the accuracy of the forecast significantly impacts the quality of operation plans. Appropriate (accurate) forecasts are important to production planning and inventory management (Lee & Adam, Jr, 1986). Inventory forecasting becomes an essential factor to support business decision making. Hogarth and Makridakis (1981) assessed forecasting accuracy and planning effectiveness in organizations and provided guidelines to calibrate expectations (Hogarth & Makridakis, 1981). Fildes (1979) pointed out that two questions dominate any assessments of a forecasting procedure (Fildes, 1979; Silver et al., 1998).

> "First, are the results statistically satisfactory? Question (a)
>
> Second, will the procedure, once developed, be used and perform in a cost-effective fashion?" Question (b)

While many researches in statistics and inventory management have been concentrated on the former question, it is the latter that is often of more concern to the practising forecaster. In this section, we review inventory forecasting evaluation metrics from statistical and managerial views.

2.3.1 Statistical evaluation metrics

While time series prediction models predict the future value based on the historical data set, measurements on reliability and accuracy of prediction models are indicators to define how well the models fit in the time series (Silver et al., 1998). To answer question (a) in section 2.3, many statistical measurements are reviewed for finding out the optimal solution from many forecasting models (Silver et al., 1998). Suppose that we have two types of information for each time period, specifically, the actual observed demands, X_1, X_2, ..., X_n and the predicted demands, $x_{0,1}$, $x_{1,2}$, ..., $x_{n-1,n}$, where $x_{t,t+1}$ is the one period ahead predicted value at period $t+1$ with predictor variables up to period t.

2.3.1.1 Mean square error (MSE)

The measure of variability is often used in fitting of squared errors of a straight line to historical data. The most commonly used indicator is the mean square error (MSE):

$$\mathrm{MSE}=\frac{1}{n}\sum_{t=1}^{n}(X_t - x_{t-1,t})^2 \qquad \text{Equation 2-20}$$

2.3.1.2 Mean absolute deviation (MAD)

An alternate measure of variability, the mean absolute deviation (MAD), was originally recommended because of its computational simplicity. However, the MAD is of less practical importance with the advent of computers. Nevertheless, it is still intuitive and is illustrated in Equation 2-21.

$$\mathrm{MAD}=\frac{1}{n}\sum_{t=1}^{n}| X_t - x_{t-1,t}| \qquad \text{Equation 2-21}$$

2.3.1.3 Mean absolute percentage error (MAPE)

The mean absolute percentage error (MAPE) is another intuitive measurement of variability. In general, because it is expressed as a percentage, it is not affected by the magnitude of the demand values. The expression can be illustrated in Equation 2-22. However, it is not appropriate if demand values are very low.

$$\mathrm{MAPE}=[\frac{1}{n}\sum_{t=1}^{n}| X_t - x_{t-1,t}|/X_t] \times 100 \qquad \text{Equation 2-22}$$

2.3.2 Managerial evaluation metrics

Time series forecasting is now a prerequisite to inventory decisions in practice (Gardner, 1990). However, statistical measures of forecast variability may not always be good indicators for decision makers.

"Forecasting is not done for its own sake; it is meaningful only as it relates to and supports decision making within the production system. Hence, managerial considerations regarding cost, effectiveness, and appropriateness of the forecasting function must be a part of the domain of forecasting."

Words above, from Blagg et al. (1980) on the study guide for the Certification Program of the American Production and Inventory Control Society, emphasized the evaluations from managerial review in terms of profit maximization and cost minimization (Silver et al., 1998). In 1990, Gardner used Equation 2-23 to minimize the sum of total variable costs in inventory control system (Silver, 1981; Gardner, 1990):

$$TVC = A\,(4D/Q) + IC\,(Q/2 + R - LD + B) + SE\,(B/U) \qquad \text{Equation 2-23}$$

where:

A = administrative cost of placing an order on procurement plus the manufacturer's production set-up cost.

D = quarterly demand forecast.

Q = order quantity.

I = inventory holding cost rate including storage, obsolescence, and opportunity costs.

C = unit purchase cost of the item.

R = reorder point, composed of lead-time demand plus safety stock.

L = procurement lead-time expressed in number of quarters.

B = expected number of units of stock backordered at any random point in time.

S = shortage cost per customer requisition backordered.

E = essentiality code for the inventory item.

U = number of units of stock per customer requisition.

$A\,(4D/Q)$ = replenishment cost, the expected ordering cost for one year.

$IC\ (Q/2 + R - LD + B)$ = Carrying costs of number of stocks on hand at any random point in time.

$SE\ (B/U)$ =shortage cost in short run.

A few studies are available on the interactions between forecasting and inventory decisions. Many managerial evaluation metrics are reasonably developed to evaluate time series forecast results. In 1990, E. S. Gardner performed a forecasting experiment on a data set that captures daily transactions of 50,000 inventory items over nine years (Gardner, 1990). Due to unavailability and low accuracy of alternative statistical measurements, trade-off curves between investment and customer service for each forecasting model were developed to support decision making in inventory management. In 1988, Bodily and Freeland designed a theoretical data set to evaluate demand and booking forecasting results among several forecasting models (Bodily & Freeland, 1988).

Chapter 3

Study and experimental design

Most, if not all, retailers collect and manage their business operating information in powerful database systems. Discovery of trends and patterns from data sets attracts tremendous attention from researchers. It is considered an essential approach to support business decisions, minimize costs, optimize profits and improve business performance. The goal of this study is to use data mining techniques to support business decisions and improve business performance. More specifically, we will use time series clustering and time series prediction techniques to forecast future quantity demand of each product in inventory management systems, so that business operating costs will be minimized and profit will be optimized simultaneously.

3.1 Overview of a small retail chain of specialty stores

The experimental data set is provided by an independent small retail chain of specialty stores. In this section, we will introduce the nature of this business and the available data set. Overviews of business operations together with a summary of sales distribution are included as well.

3.1.1 Nature of the business

The retail chain owns three stores in Toronto. Similar to many local grocery stores, it serves regular groceries and specialty products which are difficult to find elsewhere.

Most customers are from the neighbouring area. A few customers travel significant distances to get specialty products from the store. Weekly flyers and the quality of products keep attracting customers. Online shopping and E-flyers are some of the directions of development for future business.

3.1.2 Available data

The retailer collects and manages business operations records in a database management system. As a retail chain, the data set captures information on customers, products, suppliers, and business operations from January 2005 to September 2007. Product price, cost, sales quantity, and profits are some of the important factors we will focus on in this study.

3.1.3 Overviews of business operations

Compared to multinational retailers, like Wal-Mart, the store's revenue is relatively small. However, many customers are loyal to the store and thousands of products are sold successfully there. According to the data set, more than 177,000 sales transactions are made in 33 months. Sales revenue, number of products, and number of customers are included in an overview of annual business operations as shown in Table 3-1. In total, there are 20,812 distinct customers, 10,841different products and $8,737,000 sales revenue over two years and nine months. Similarly, an overview of quarterly business operations is shown in Table 3-2.

Attribute	2005	2006	2007(9months)	Total
Sales revenue	$1,574,079	$3,885,394	$3,278,189	$8,737,662
Number of products	5782	7567	7587	10841
Number of customers	3824	9818	9371	20812

Table 3-1: An overview of annual business operations

Attribute	2005 Quarter 1	2005 Quarter 2	2005 Quarter 3	2005 Quarter 4	2006 Quarter 1	2006 Quarter 2	2006 Quarter 3	2006 Quarter 4
Sales revenue	$387,973	$368,857	$369,781	$447,468	$1,010,460	$891,598	$955,130	$1,028,206
Number of products	3502	3472	3551	3881	5368	4908	4883	5127
Number of customers	1585	1494	1789	1966	4670	4582	4901	5051

Attribute	2007 Quarter 1	2007 Quarter 2	2007 Quarter 3
Sales revenue	$1,033,560	$1,048,181	$1,196,448
Number of products	5274	5505	5876
Number of customers	5383	5369	5347

Table 3-2: An overview of quarterly business operations

In this study, our objective is to improve inventory management. That is, we aim to find out sales patterns of products in order to properly predict future quantity demand. Sales quantity, product selling price and cost are some of the important indicators that interest inventory managers. Since data in 2007 only includes business records for 9 months and data in 2006 includes more records than data in 2005, our analysis emphasizes products sold in 2006. Table 3-3 illustrates annual sales quantity and profit distributions of products. For each product, annual sales quantities and profits are analyzed. In 2006, the minimum sales quantity (Min) is -3, and maximum sales quantity (Max) is 3825. In addition, the mean sales quantity of products is 23.67 and the standard deviation of sales quantity is 77.04. The minimum, maximum, mean value and standard deviation of annual profits are $-2985, $21882, $213.14 and 660.71. Similarly, Tables 3-4, 3-5, 3-6 and 3-7 illustrate quarterly and monthly sales quantity and profits distributions of products in 33 months.

	Attribute	2005	2006	2007(9months)
Quantity	Min	-2	-3	-6
	Max	686	3825	2650
	Mean	12.24	23.67	20.13
	Standard Deviation	27.14	77.04	59.61
Profit	Min	$-1647	$-2985	$-22935
	Max	$7059	$21882	$25870
	Mean	$115.11	$214.14	$179.10
	Standard Deviation	296.71	660.71	655.21

Table 3-3: An annual overview of products

		2005				2006			
	Attribute	Quarter1	Quarter2	Quarter3	Quarter4	Quarter1	Quarter2	Quarter3	Quarter4
Quantity	Min	-1	-3	-3	-1	-4	-1	-2	-4
	Max	239	127	107	230	1015	850	1077	883
	Mean	3.02	2.88	2.87	3.47	6.23	5.42	5.80	6.21
	Standard Deviation	7.88	6.67	6.57	8.50	21.82	17.39	20.82	20.06
Profit	Min	$-65	$-1843	$-63	$-84	$-170	$-79	$-6642	$-119
	Max	$2,026	$1,640	$1,381	$2,114	$6,439	$5,414	$4,842	$5,187
	Mean	$28.46	$27.09	$27.78	$31.79	$55.92	$50.1	$50.8	$57.33
	Standard Deviation	78.53	81.04	72.65	85.69	184.1	160.2	184.5	175.7

Table 3-4a: A quarterly overview of products (2005 and 2006)

		2007		
	Attribute	Quarter1	Quarter2	Quarter3
Quantity	Min	-2	-2	-6
	Max	573	939	1138
	Mean	6.30	6.42	7.42
	Standard Deviation	19.01	19.27	23.75
Profit	Min	$-1655	$-22952	$-209
	Max	$25,478	$3,818	$5,192
	Mean	$59.11	$55.41	$64.58
	Standard Deviation	342.5	314.41	190

Table 3-4b: A quarterly overview of products (2007)

2005	Attribute	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Quantity	Min	-2	-2	-3	-3	-2	-3	-1	-5	-2	-1	-1	-2
	Max	49	87	134	59	48	43	57	53	50	112	58	84
	Mean	0.99	0.98	1.06	0.98	1.03	0.87	0.99	0.96	0.91	1.06	0.99	1.42
	Standard Deviation	2.52	3.02	3.54	2.68	2.57	2.27	2.53	2.52	2.32	3.18	2.59	3.92
Profit	Min	$-21	$-17	$-71	$-1888	$-276	$-42	$-63	$-66	$-58	$-81	$-23	$-86
	Max	$690	$648	$688	$644	$644	$595	$493	$687	$560	$512	$711	$1,077
	Mean	$9.61	$9.19	$9.66	$8.79	$9.78	$8.52	$9.49	$9.3	$8.98	$9.67	$9.34	$12.8
	Standard Deviation	27.5	28	30.1	37.3	29.7	26.3	27.1	26.8	26	27.8	28	38.2

Table 3-5: A monthly overview of products sold in 2005

2006	Attribute	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Quantity	Min	-4	-2	-4	-3	-1	-1	-1	-2	-2	-3	-4	-1
	Max	356	457	300	205	407	238	415	346	316	357	252	326
	Mean	2.73	1.6	1.9	1.76	1.83	1.83	2.12	1.8	1.89	1.95	1.99	2.26
	Standard Deviation	9.69	7.12	6.64	5.56	6.98	5.79	9.3	6.45	6.58	7.05	6.03	8.34
Profit	Min	$-35	$-43	$-93	$-86	$-71	$-31	$-101	$-136	$-7,657	$-47	$-65	$-145
	Max	$3,375	$1,461	$1,603	$1,656	$1,646	$2,112	$1,715	$1,771	$3,295	$1,510	$1,725	$1,952
	Mean	$24.1	$14.4	$17.4	$16.2	$16.7	$17.2	$17.4	$16.5	$16.87	$18.3	$18.8	$20.2
	Standard Deviation	83.9	49.7	58.7	54.7	56.1	58.3	59.3	55	111.3	59.2	58.5	66.1

Table 3-6: A monthly overview of products sold in 2006

2007	Attribute	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.
Quantity	Min	-1	-3	-2	-2	-2	-2	-1	-6	-5
	Max	284	202	249	323	352	264	395	382	361
	Mean	2.325	1.893	2.081	2.091	2.178	2.153	2.503	2.121	2.783
	Standard Deviation	8.16	5.86	6.42	6.68	6.95	6.58	9.62	7.30	8.55
Profit	Min	$-122	$-1820	$-179	$-85	$-22965	$-88	$-82	$-99	$-127
	Max	$2,005	$25,426	$1,443	$1,446	$1,711	$1,264	$1,278	$1,594	$2,320
	Mean	$19.66	$20.77	$18.68	$19.44	$16.67	$19.29	$21.05	$18.78	$24.76
	Standard Deviation	67.05	298.9	58.78	60.79	270.7	57.94	65.21	58.16	77.46

Table 3-7: A monthly overview of products sold in 2007

Some products are sold throughout a whole year, while others are sold seasonally. Quarterly distributions of products sold in 2005 and 2006 are illustrated in Tables 3-8 and 3-9. The shaded diagonals indicate the number of products sold in each quarter. For example, in 2006, 5314 products are sold in the first quarter and 4879 products are sold in the second quarter. The other cells indicate the intersections of the number of products sold in two quarters. For example, in 2006, 3821 products are sold in the first quarter and the second quarter and 3700 products are sold in the first quarter and the third quarter. Similarly, monthly distributions of products sold in 2005 and 2006 are illustrated in Tables 3-10 and 3-11.

2005	Quarter1	Quarter2	Quarter3	Quarter4
Quarter1	**3470**	2483	2335	2351
Quarter2	2483	**3427**	2481	2457
Quarter3	2335	2481	**3508**	2601
Quarter4	2351	2457	2601	**3841**

Table 3-8: The quarterly distribution of products sold in 2005

2006	Quarter1	Quarter2	Quarter3	Quarter4
Quarter1	**5314**	3821	3700	3700
Quarter2	3821	**4879**	3785	3695
Quarter3	3700	3785	**4847**	3824
Quarter4	3700	3695	3824	**5085**

Table 3-9: The quarterly distribution of products sold in 2006

2005	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Jan.	**2118**	1244	1250	1217	1240	1080	1164	1119	1096	1110	1084	1208
Feb.	1244	**2084**	1209	1232	1234	1084	1134	1152	1080	1088	1063	1190
Mar.	1250	1209	**2079**	1239	1285	1139	1185	1173	1120	1126	1084	1233
Apr.	1217	1232	1239	**2100**	1319	1173	1209	1209	1147	1159	1102	1243
May.	1240	1234	1285	1319	**2195**	1226	1289	1280	1191	1223	1195	1292
Jun.	1080	1084	1139	1173	1226	**1961**	1178	1184	1104	1083	1097	1193
Jul.	1164	1134	1185	1209	1289	1178	**2177**	1276	1258	1216	1185	1295
Aug.	1119	1152	1173	1209	1280	1184	1276	**2138**	1241	1247	1244	1333
Sep.	1096	1080	1120	1147	1191	1104	1258	1241	**2063**	1211	1203	1279
Oct.	1110	1088	1126	1159	1223	1083	1216	1247	1211	**2132**	1253	1355
Nov.	1084	1063	1084	1102	1195	1097	1185	1244	1203	1253	**2121**	1359
Dec.	1208	1190	1233	1243	1292	1193	1295	1333	1279	1355	1359	**2589**

Table 3-10: The monthly distribution of products sold in 2005

2006	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Jan.	**4056**	2259	2360	2312	2301	2304	2290	2231	2283	2284	2314	2351
Feb.	2259	**2991**	2128	2065	2038	1990	1951	1899	1913	1921	1946	1954
Mar.	2360	2128	**3199**	2197	2165	2138	2052	2037	2054	2042	2069	2076
Apr.	2312	2065	2197	**3165**	2159	2128	2077	2026	2052	2041	2042	2063
May.	2301	2038	2165	2159	**3200**	2225	2202	2146	2109	2132	2127	2139
Jun.	2304	1990	2138	2128	2225	**3304**	2241	2221	2178	2131	2177	2183
Jul.	2290	1951	2052	2077	2202	2241	**3216**	2208	2200	2161	2174	2176
Aug.	2231	1899	2037	2026	2146	2221	2208	**3199**	2212	2187	2211	2184
Sep.	2283	1913	2054	2052	2109	2178	2200	2212	**3277**	2270	2281	2289
Oct.	2284	1921	2042	2041	2132	2131	2161	2187	2270	**3312**	2357	2354
Nov.	2314	1946	2069	2042	2127	2177	2174	2211	2281	2357	**3472**	2471
Dec.	2351	1954	2076	2063	2139	2183	2176	2184	2289	2354	2471	**3557**

Table 3-11: The monthly distribution of products sold in 2006

From the Tables above, it is not difficult to discover that some products are mostly sold in certain month(s) or quarter(s). Products are considered as single period selling products if their selling ratios in one period are high. Here, the ratio is calculated as Equation 3-1:

$$\frac{pk}{P-pk} \qquad \text{Equation 3-1}$$

Where *pk* is the sales quantity in period *k, P* is the total sales quantity throughout the year. The ratio is considered high if its value is greater than 10. Table 3-12 illustrates the distribution of single quarter selling products. For instance, in 2006, 800 products were mostly sold in the first quarter, 336 products are mostly sold in the second quarter, etc. Table 3-13 illustrates the distribution of single month selling products.

	Quarter1	Quarter2	Quarter3	Quarter4
Number of products (2005)	478	292	332	662
Number of products (2006)	800	336	306	579

Table 3-12: The distribution of single quarter selling products

Month	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Number of products (2005)	132	156	106	79	96	74	104	92	91	108	115	300
Number of products (2006)	491	93	99	110	76	105	85	91	81	92	127	188

Table 3-13: The distribution of single month selling products

Sales quantity and profit are not normally distributed among products. Annual sales quantities of 7568 products for 2006 are sorted in an ascending order to study ranked distribution of products. To give a clear overview of majority sales quantity distribution, 76 products, which are 1% of total products, are dropped from the distribution table since they are considered as outliers. That is, 38 products with the lowest sale quantities and another 38 products with the highest sales quantities are omitted. Annual sales quantities of 7498 (99%) products are distributed in Figure 3-1, the rank distribution of annual sales quantities. We can see that most products are sold less than 50 times in 2006 and these include 6732 (89.78%) products. Similarly, Figure 3-2 illustrates the rank distribution of annual profits in 2006.

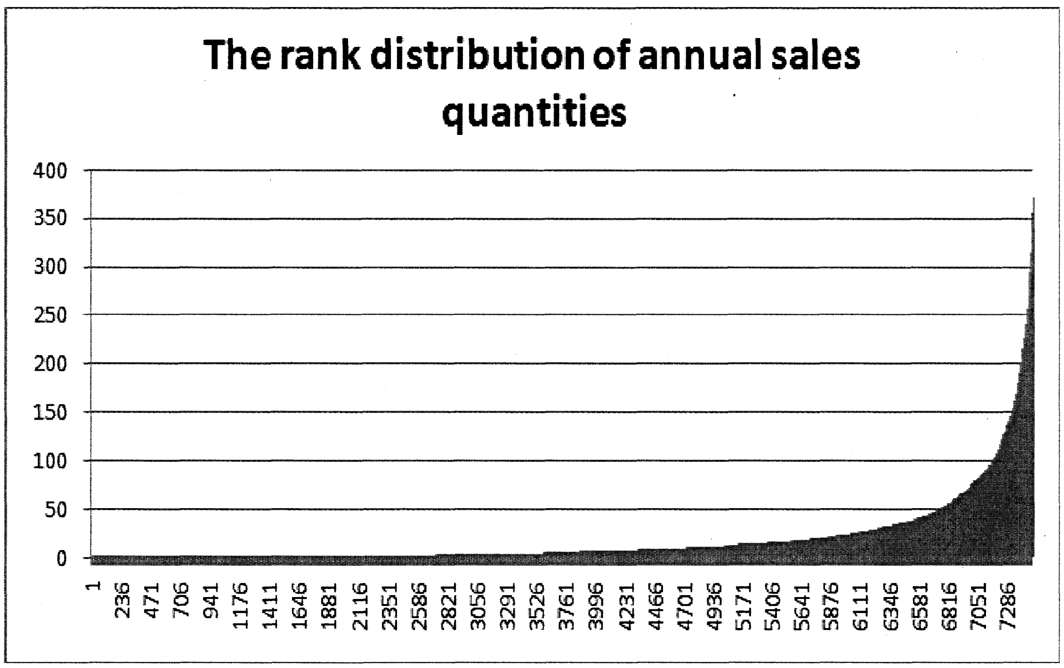


Figure 3-1: The rank distribution of annual sales quantities in 2006

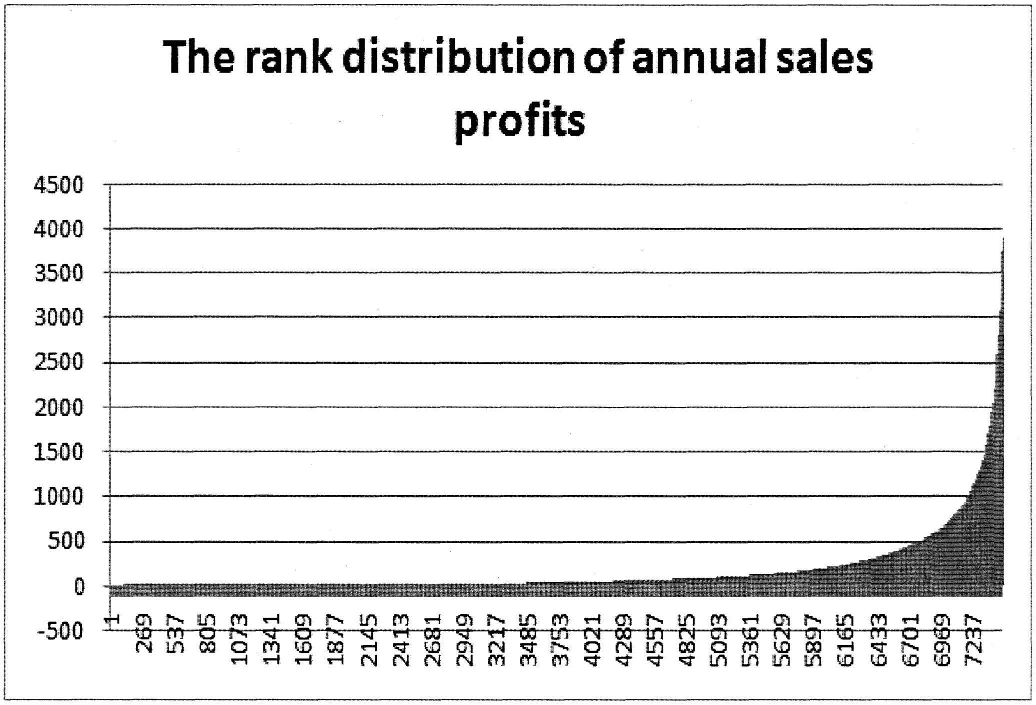


Figure 3-2: The rank distribution of annual sales profits in 2006

In addition, a frequency distribution of annual sales quantities is shown in Figure 3-3. The figure shows that in 2006, 5698 products were sold less than 20 times, 834 products were sold from 20 to 39 times, etc. Similarly, Figure 3-4 shows the frequency distribution of annual sales profits in 2006.

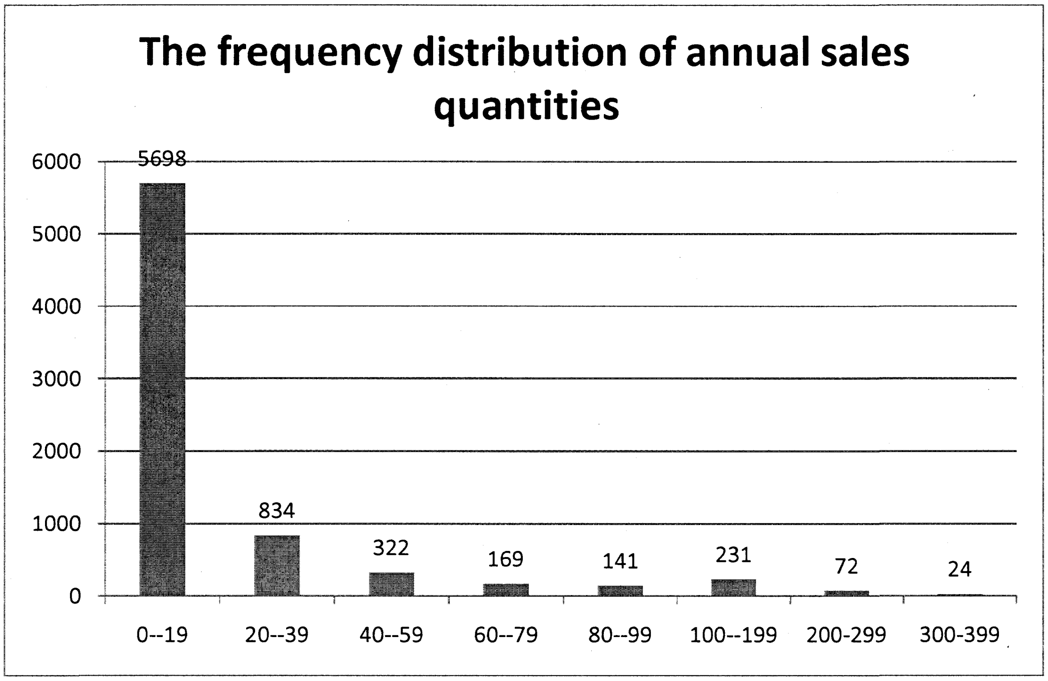


Figure 3-3: The frequency distribution of annual sales quantities in 2006

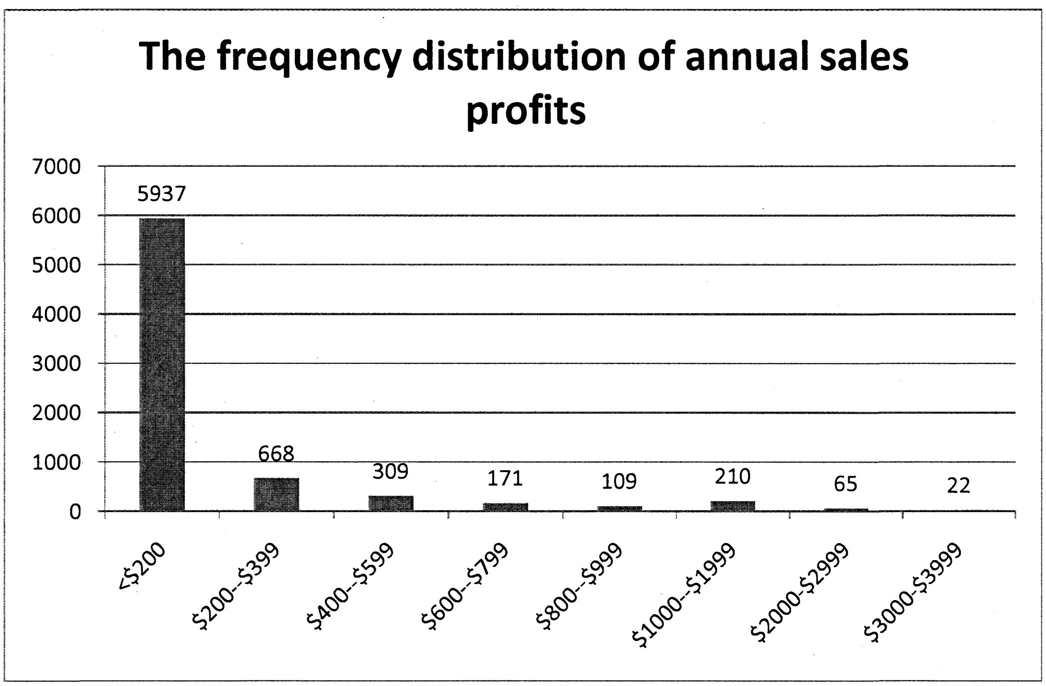


Figure 3-4: The frequency distribution of annual sales profits in 2006

3.1.4 Summary of sales distribution

In general, the retail chain operates relatively well. The revenue, number of products and customers are recognizable. Huge differences between minimum and maximum sales quantities and profits are illustrated in annual, quarterly and monthly overviews. Thus, products have their own sales patterns. Moreover, some products may share the same patterns. Seasonal selling distributions are clearly shown for some products. Categorizing products, based on their sales patterns, may ease inventory management. However, some data occurrences do not make sense. For example, most of the minimum sales quantities are negative. This may be caused by records of product return and/or errors of original data entry. Data study and data cleaning are some of the future tasks required for empirical experiments.

3.2 Data preparation

Real world data is collected and managed to record business operations. To meet the requirements of data mining techniques, real world data needs to be prepared through many processes, such as, data collection, data study, data cleaning, data extraction, data transformation and data consolidation. Some of these processes will be discussed in this section.

3.2.1 Business understanding and data study

A comprehensive understanding of business operations and performance provides data miners with clear definitions of data mining goals. It also accelerates data understanding and preparation processes, such as, data cleaning, extraction and consolidation.

A huge amount of business data regarding customers, products, suppliers and business operations is collected in the experimental data set. Similar to many business companies, the data set is constructed in Relational Database Management System (RDMS). RDMS is based on the relational model that was introduced by E.F. Codd in 1970 (Codd, 1970). It is perhaps the most popular database model that is used commercially. The original data set can be viewed in Microsoft Access. Figures 3-5a and 3-5b illustrates the database schema of business operations data. Tables *Product, Invoice, Customer, Supplier* and *Employee* in the Figure collect and record details of business operations.

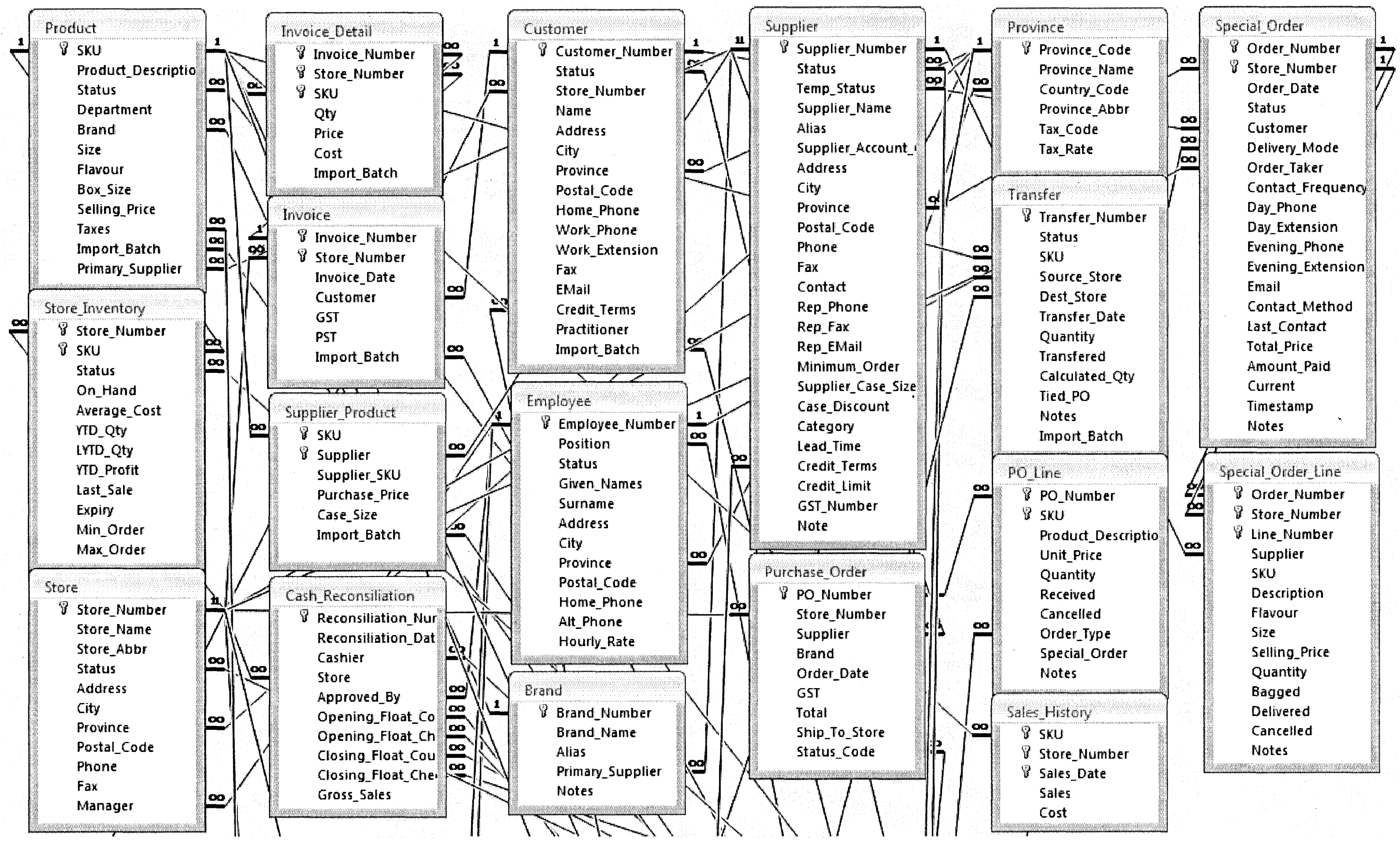


Figure 3-5a: The database schema of the business data set

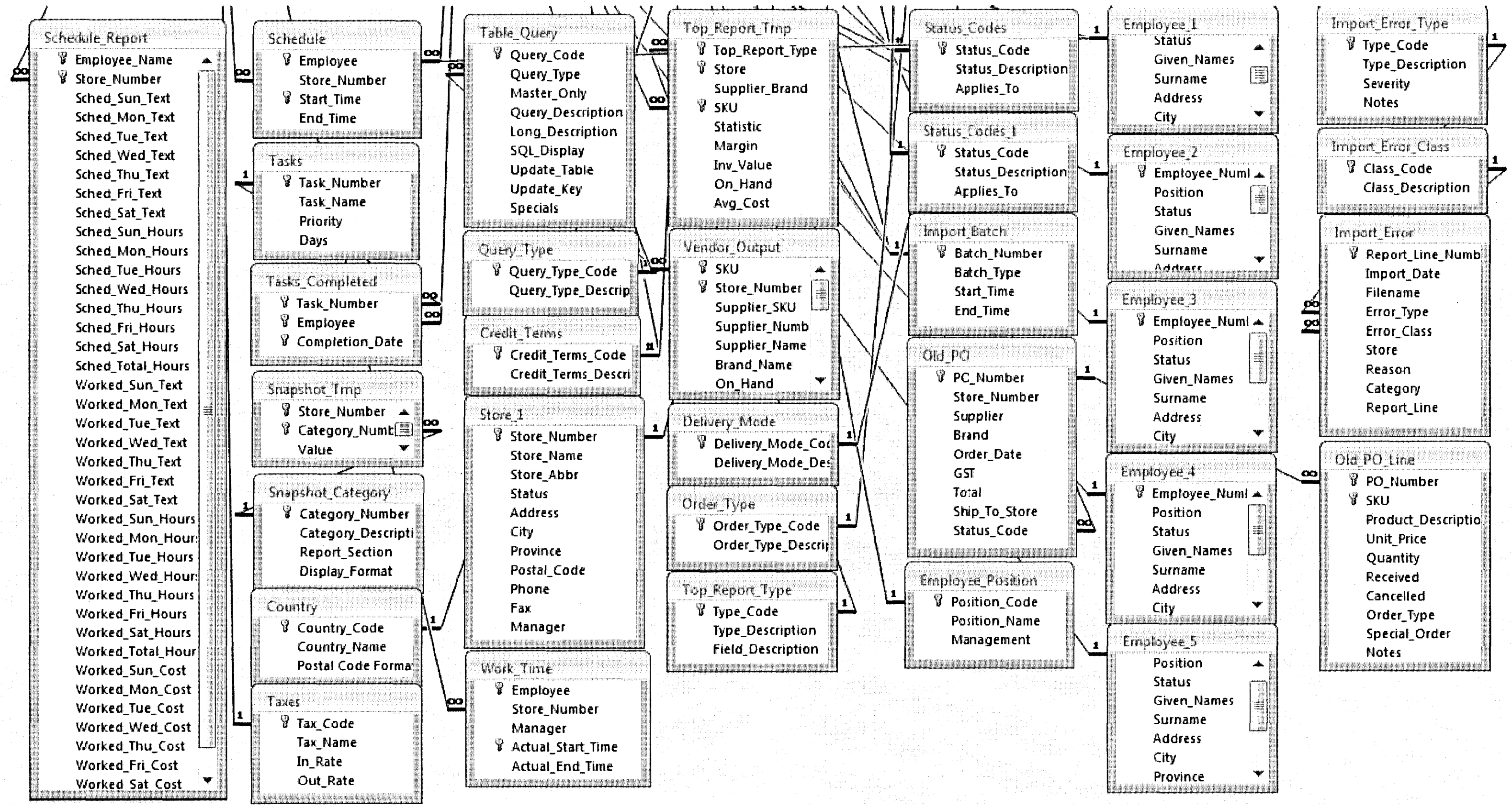


Figure 3-5b: The database schema of the business data set

For experimental purposes, the data set is migrated to a MySQL database. Further data preparation processes are conducted on the data set through MySQL database. Due to the incompleteness, inconsistency and noisiness of the original data set, we cleaned the problematic data before performing any data mining experiments.

3.2.2 Data cleaning and extraction

Data cleaning is a critical process in data mining projects. It removes incomplete, inconsistent, and noisy data and assures a high quality of experimental data. In this study, we focus on discovering the sales patterns of products within a fiscal year. That is, only products with positive annual sales quantities are eligible for the experiments. On the other hand, products are considered as invalid samples if their annual sales quantities are negative or there are no sales records associated with them. This could be caused by records of product returns and incomplete and inconsistent data entries. Some invalid product samples in 2005 and 2006 are shown in Table 3-14 and 3-15, respectively. For example, in 2005, product 068958048215 has a negative annual sales quantity of -1. We analyze all the sales records associated with this product, as shown in Table 3-16. Negative sales quantities may be caused by incorrect data entry or product return. The negative sales quantity on 4/28/2005 is not a record of product return since there are no sales records before it. In this case, there are no valid explanations to support this product sold -1 time on 4/28/2005 or in 2005. Therefore, the data entry for this product is either incomplete or incorrect. As a result, product 068958048215 is confirmed as an invalid experimental sample. The same verification process is applied to all invalid product samples to confirm their exclusion from this study. All the confirmed problematic product samples are omitted from this study. In the next step, only valid product data is extracted to facilitate this study.

Product ID	Annual sales quantity
-1574973	0
021718500491	0
028367829676	-1
036923001565	0
036923101005	0
036923282148	0
0682-534	0
068958021058	0
068958024042	0
068958035550	0
068958035901	0
068958048215	-1
076280924008	0

Table 3-14: Invalid product samples in 2005

Product ID	Annual sales quantity
020855704519	0
027434001113	-1
030985004502	0
062767021810	0
068958048215	0
075941001102	0
81738402229	0
88395051517	0
088395070501	-1
3.10539E+11	0
5401-2	-1
6.00726E+11	0
6.45947E+11	0
6.5801E+11	0

Table 3-15: Invalid product samples in 2006

Receipt Number	Date	Customer ID	Product ID	Quantity	Price	Cost
121533	9/18/2006	106544	068958048215	-1	-22	-13
123940	10/17/2006	0	068958048215	1	25	13
134548	4/28/2005	108352	068958048215	-1	-22	-13
134548	2/15/2007	106544	068958048215	-1	-22	-13

Table 3-16: Sales records associated with product 068958048215

Data mining projects and tasks may require different data. Inventory data is central to the present study. Time series clustering, time series prediction and simulation are some of the data mining tasks that will be performed against the target data set. Task oriented data needs to be extracted separately. Particularly, product names and weekly, monthly or quarterly sales quantities of each product are extracted to facilitate three-level time series clustering and time series prediction. In addition, product name and weekly, monthly or quarterly sales quantity, selling price and cost of each product are extracted to facilitate simulation programs.

3.2.3 Data consolidation

Task oriented data may not always be formatted as required by data mining tasks. Data consolidation processes transform data into the required formats. In this study, time series data is required for two data mining tasks: time series clustering and time series prediction. The extracted data is consolidated into time series format, such that, the data point sequence represents the same type of information measured at successive times, spaced at uniform time intervals. Most product IDs are named with numbers. However, there are a few of them named with texts. In a CSV file, large numbers may be shown in scientific notation format. This could cause confusion of product IDs recognition. Some examples of this confusion are shown in Table 3-17. Product IDs in the shaded cells are all shown as 3.71401E+11. However, they are actually 371400513019, 371400515013, 371400701010 and 371401007609, respectively.

Product ID	Quarter1	Quarter2	Quarter3	Quarter4
3.58286E+11	0	0	0	7
3.58287E+11	0	1	5	1
3601-1	0	0	1	0
3.66107E+11	2	0	3	4
3.66492E+11	2	7	3	1
3.71401E+11	1	0	0	0
3.71401E+11	1	0	0	0
3.71401E+11	0	0	2	4
3.71401E+11	0	0	1	0
3.76009E+12	0	2	0	0
400	12	12	24	17
4.00163E+12	1	0	1	1

Table 3-17: Confusion of product IDs recognition in a CSV file

In addition, product IDs are shown incorrectly if they are numbers and start with "0". Here, a CSV source file is shown below:

Product ID, Quarter1, Quarter2, Quarter3, Quarter4

021718500125, 4, 6, 2, 4

021718500132, 2, 3, 4, 7

021718500149, 4, 5, 2, 6

021718500156, 0, 0, 3, 2

021718500163, 6, 4, 4, 4

021718500170, 10, 9, 7, 3

When viewing this table in a spreadsheet, "0"s in front of product IDs are omitted, as shown in Table 3-18. This could cause some recognition problems. To eliminate this confusion, we add a letter "P" in front of every product ID to make it into text format. After this, all product IDs are shown in the text format in CSV files. Table 3-19 illustrates the consolidated quarterly sales quantity data for time series clustering and time series prediction. The previous confusion is eliminated as shown in the shaded cells. Similar consolidation processes are also applied to extracted data for simulation programs. The consolidated data can be viewed and recognized clearly in multiple formats including CSV, XLSX, and TXT formats.

Product ID	Quarter1	Quarter2	Quarter3	Quarter4
21718500125	4	6	2	4
21718500132	2	3	4	7
21718500149	4	5	2	6
21718500156	0	0	3	2
21718500163	6	4	4	4
21718500170	10	9	7	3

Table 3-18: Confusion of product IDs recognition in a CSV file

Product ID	Quarter1	Quarter2	Quarter3	Quarter4
P358286206013	0	0	0	7
P358286700016	0	1	5	1
P3601-1	0	0	1	0
P366106502412	2	0	3	4
P366492001025	2	7	3	1
P371400513019	1	0	0	0
P371400515013	1	0	0	0
P371400701010	0	0	2	4
P371401007609	0	0	1	0
P3760087360165	0	2	0	0
P400	12	12	24	17
P021718500125	4	6	2	4
P021718500132	2	3	4	7
P021718500149	4	5	2	6
P021718500156	0	0	3	2
P021718500163	6	4	4	4
P021718500170	10	9	7	3

Table 3-19: Consolidated quarterly sales quantity data for time series clustering and time series prediction.

3.3 Data mining techniques used in inventory management

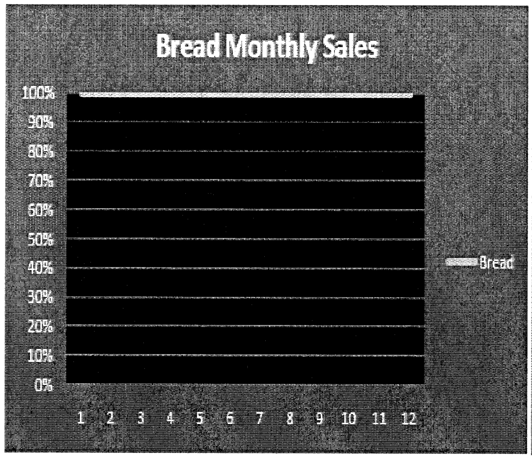
Many data mining techniques, such as, clustering, classification, prediction and association are available to discover patterns and business knowledge from a data set. In this study, time series clustering is applied to categorize products into reasonable groups based on their sales patterns. Furthermore, time series prediction is conducted to forecast quantity demand of each product, which is used to support business decisions in inventory management system.

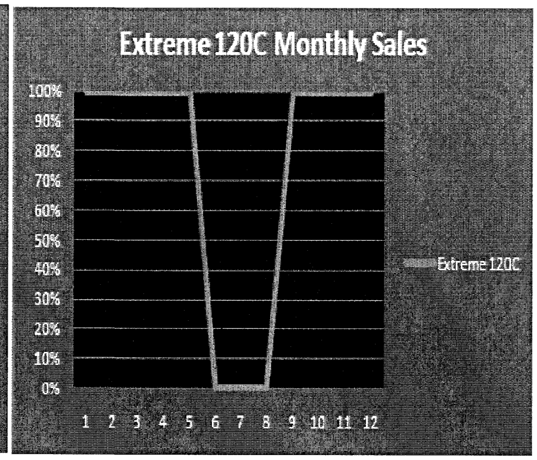
3.3.1 Products profiling and time series clustering

Time series clustering is a simple technique to categorize products into several reasonable groups based on similarity among their distributions and demand patterns. K-Means and EM, commonly used time series clustering algorithms, will be applied to categorize products in this study.

3.3.1.1 Stability analysis

Many quantity analyses will be performed based on historical sales records. Products may be distinguished into groups based on their volatilities in sales quantities. We will attempt to categorize products based on their stabilities of sales quantity in three levels: weekly, monthly and quarterly. K-Means and EM algorithms will be applied to the three level analyses. The possible categorizing results are weekly stable products, monthly stable products, quarterly stable products, and unstable products. Ideally, weekly stable products are some of the products in the group of monthly stable products. Moreover, monthly stable products are some of the products in the group of quarterly stable products. Unstable products are products that have large volatilities in weekly, monthly and quarterly sales quantity. Figure 3-6 illustrates a sample result of stability analysis of products: Bread and Extreme 120C. We may also compare stability analysis results among different years to ensure confidence on the stability of products.





Stable monthly product- Bread Unstable monthly product –Extreme 120C

Figure 3-6: Sample stability analysis result of Bread and Extreme 120C

3.3.1.2 Seasonality analysis

Within the group of unstable products, seasonal patterns provided us another criterion to further categorize products. That is, we will categorize unstable products into groups based on their monthly and quarterly sales patterns. K-Means and EM clustering algorithms will be applied again to clustering products. Clustering results may provide us reasonable groups: such as, spring, summer, fall, winter, spring-fall and winter-spring products. However, sales-seasons may be defined based on geographic location of retail companies and sales power of products instead of the traditional way, where a quarter includes three months. For example, the sales-season for Christmas products is December, since most of their transactions are made in December to celebrate Christmas. In addition, we may perform stability analysis on each group to find out their stabilities in sales-season and off-season. This will help us indentify predictor variables and predictable periods in time series prediction. Above all, we may categorize products into following groups:

- Stable weekly products
- Stable monthly products
- Stable quarterly products

The following groups apply to four seasons:

- Stable weekly in sales-season and stable weekly in off-season products
- Stable weekly in sales-season and stable monthly in off-season products
- Stable weekly in sale-season and stable quarterly in off-season products
- Stable weekly in sales-season and unstable in off-season products
- Stable monthly in sales-season and stable weekly in off-season products
- Stable monthly in sales-season and stable monthly in off-season products
- Stable monthly in sales-season and stable quarterly in off-season products
- Stable monthly in sales-season and unstable in off-season products
- Stable quarterly in sales-season and stable weekly in off-season products
- Stable quarterly in sales-season and stable monthly in off-season products
- Stable quarterly in sales-season and stable quarterly in off-season products
- Stable quarterly in sales-season and unstable in off-season products
- Unstable in sales-season and stable weekly in off-season products
- Unstable in sales-season and stable monthly in off-season products

- Unstable in sales-season and stable quarterly in off-season products
- Unstable in all seasons products

3.3.2 Inventory forecasting and time series prediction

Time series prediction, which predicts future values of a time series, plays a critical role in forecasting quantity demand in business operations. Regression analysis, neural networks, exponential smoothing and autoregressive integrated moving mean (ARIMA) are some of the widely used time series prediction techniques in inventory management. We will apply some of the popular prediction models on our dataset. SPSS, the statistical software supported by IBM, is known as one of the most powerful data analysis and time series prediction application. For any given volatile time series, seasonal prediction techniques are able to find seasonal patterns and predict future values respectively. However, the accuracy of prediction may be reduced due to different volatilities in sales-seasons and off-seasons. We will compare prediction results from SPSS prediction models with prediction results from our forecasting models.

3.3.2.1 Inventory forecasting models

Time series prediction techniques, predictable periods and predictor variables are essential factors of forecasting in inventory management. Our proposed inventory forecasting steps are:

Step 1: Select a time series prediction technique.

Step 2: Identify predictable periods of the given product.

Step 3: Identify predictor variables of the given product.

Step 4: Forecast the future values with selected techniques and variables.

We will create forecasting models based on the steps above. First, one of the commonly used prediction techniques will be selected (Step1). For a given product, we will identify the predictable periods and predictor variables based on stability and seasonality analyses and requirements of the prediction technique (Step 2 and 3). Then, we will predict future values in the predictable periods with the selected prediction techniques and predictor variables. Many different forecasting models will be created by repeating those steps with different prediction techniques for the same product. Prediction results will be evaluated and compared in sections 3.4 and 3.5 to find out the best fit prediction model for each product.

3.4 Inventory forecasting evaluation

The distance between predicted values and actual values indicates the accuracy of time series prediction model. The lower the distance is, the more accurate the time series prediction model is. As we discussed in section 2.3.1, time series prediction models are often evaluated by statistical measurements, such as, mean square error (MSE), mean absolute deviation (MAD) and mean absolute percentage error (MAPE). In this study, MAPE will be applied to evaluate inventory forecasting models. Forecasting models with the lowest errors are considered as the most reliable and accurate solutions.

3.5 Simulation of business operations and performance

The goal of inventory management is to support business operations with the least amount of products without ever running out. Statistical metrics may not always be good indicators for finding out the optimal solutions. Total cost, total profits and shortage periods are some of the critical metrics that attract managers' attention. Some managerial adjustments may be applied to satisfy this goal. Based on managerial experience, adding a certain level of prediction buffer beyond inventory prediction models, to avoid lost sales opportunity, is a typical example.

3.5.1 Inventory system operation simulation

To further support business decisions, a simulation program will be created to simulate business operations and performance based on historical sales records and prediction results from inventory forecasting models. In addition, the simulation program will generate business operation reports. In the report, the length of shortage periods and total costs will be measured based on the simulated business operations for each product.

3.5.2 Cost management

Cost management is one of the most important issues in inventory management. As previously discussed in section 2.3.2, according to Gardner's total variable cost theory of inventory control system (1990), total variable cost includes replenishment cost, carrying cost and shortage cost. Replenishment costs, which are associated with each order, change negligibly with inventory forecasting models. Carrying costs include stock-keeping costs of products against bank interests. Shortage costs are accrued from losing of customer loyalty and product profits due to unavailability of products. Carrying cost and shortage cost are highly affected by inventory forecasting models. Therefore, effective inventory forecasting solutions improve inventory management by minimizing total costs and maximizing profits simultaneously.

3.6 Experiment applications and softwares

This section describes the details of the experiments including the software packages that were used. These details include data preparation, product profiling and analyses with time series clustering, inventory forecasting with time series prediction, and business simulation.

Data preparation

The experimental data set is obtained from a small retail chain store. It is in an NDB file format. The data set is loaded into a MySQL database, so that the data can be accessed, searched and retrieved with several SQL queries. Data normalization and data consolidation was performed with Java programs.

Product profiling and analyses

Time series clustering techniques were used in product profiling and analyses. This study used Weka to perform time series clustering analyses. Weka is a powerful data mining application that contains tools for data pre-processing, clustering, classification, prediction, association, and visualization. Several clustering algorithms, such as EM and K-Means, are available in Weka. Java programs were used for data normalizations, product categorizations and comparisons.

Inventory forecasting

Time series prediction techniques were used in inventory forecasting. Several common time series clustering techniques are available in SPSS, a well-known data mining application supported by IBM. This study performed inventory forecasting with SPSS. SPSS atutomation scripts were created with Java programs. Since the amount of computation is significant, scripts were divided into several acceptable sections and fed in SPSS. In addition, Java programs were required to calculate MAPEs, compare MAPEs, identify optimal solutions, compare optimal solutions, and generate reports.

Business simulation

A business simulation program was written to simulate business operations based on historical sales quantities, prices, costs, and predicted quantity demands. It performed cost/benefit analysis and generated business reports. The simulation program was written in Java. In addition, Java programs were written to compare business reports and identify optimal solutions.

Chapter 4

Product profiling and time series clustering

In inventory management systems, products can be distinguished by brand, price, cost, size, and sales quantity. In this section, products are profiled on their volatilities and seasonalities based on sales quantity. Since the data set only captures 9 months data from 2007, stability and seasonality analyses are only performed on products sold in year 2005 and 2006. In addition, products will be categorized into reasonable groups based on their sales patterns. K-Mean and EM clustering algorithms are applied in time series clustering processes to analyze products' stabilities and seasonalities.

4.1 Stability analysis

Products are called stable if their sales quantities change negligibly over allotted periods. On the other hand, products are defined as unstable if their sales quantities change greatly with time. According to section 3.3.1.1, the stability analysis is performed in three levels: weekly, monthly and quarterly. This stability analysis is carried out from "the least" to "the greatest", that is, the experiment starts with weekly sales quantities analysis, and then follows by monthly and quarterly sales quantities analysis. Each level of analysis categorizes products into two groups: a stable and an unstable group. Moreover, stable weekly products can be considered stable monthly products, and stable monthly products can be considered stable quarterly products, and so on. Thus, we perform monthly stability analysis on unstable weekly products and quarterly stability analysis on unstable monthly products. Finally, products in unstable quarterly groups are used in seasonality analysis in section 4.2.

This study applies time series clustering techniques to categorize products into reasonable groups. However, results generated by the time series clustering may not meet the goals of analysis if the original data set is noisy. Thus, further time series clustering analysis is required to refine results. Therefore, we apply two tiers of clustering analysis on each level of stability analysis. That is, Tier-1 clustering analysis provides us with preliminary results, which indicate stable groups. Tier-2 clustering analysis is performed based on these stable groups to identify even more stable products. In Tier-1 stability analysis, we pay more attention to the value of P so that we can categorize frequently sold products into stable product groups. Since stable products that were defined in the Tier-1 clustering analysis are clustered in the Tier-2 stability analysis, we pay more attention to the value of A. Thus, products with similar sales quantities in each period are considered stable.

4.1.1 A refined stability analysis

This study considers many criteria for stability analysis. However, most of them do not work properly. For example, the result of the mean divided by standard deviation for each product is volatile. The mean value of the standard score (z-score) is reasonable for all-season products. However, products with zero sales in most time intervals are considered stable. In fact, they are really seasonal products because they are only sold in a small number of time periods. In this study, we developed a refined stability analysis. It combines two reasonable criteria: the mean of absolute z-score and the percentage of non-zero values. The mean of absolute standard score, denoted by A, is:

$$z=\frac{x-\mu}{\sigma} \qquad A=\frac{\sum|z|}{n} \qquad \text{Equation 4-1}$$

where x is the value of data object, which is the sales quantity of a given period, μ is the mean of the population, σ is the standard deviation of the population and n is the total number of data objects. The value of A indicates how stable the product is based on periodical sales quantities. The lower the value of A, the more stable the product.

The percentage of non-zero values, denoted by P, is:

$$P=\frac{m}{n} \qquad \text{Equation 4-2}$$

where m is the number of data objects with values not equal to zero. The value of P indicates the frequency of the product sold in equivalent periods. Therefore, the range of value P is from 0 to 1. Zero means that the product has no sales records in any period. One means that the product has been sold in every period. Products with higher values of P are considered to be sold more often, which makes them more stable.

Data sets are usually normalized before any data mining tasks. There are two popular methods for normalization: max-normalization, which divides object values by the maximum value, and mean-normalization, which divides objects values by the mean value. Here, we apply max-normalization method on experimental data sets.

Table 4-1 shows a sample of time series clustering data sets. For example, P128 has a higher value of P compared to P114. That is, P128 has been sold more often than P114. In addition, the A value of P128, 0.870237, is lower than the A value of P114, which is 0.913061. Therefore, P128 is more stable than P114.

Product ID	Mean of absolute z-score (A)	Percentage of non-zero values (P)
P061998079829	0.905559	0.980769
P061998084458	0.940804	1
P068958011219	0.821932	0.884616
P107	0.84088	0.903846
P114	**0.913061**	**0.903846**
P128	**0.870237**	**1**
P408	0.889739	0.980769
P418 120	0.886479	0.961538
P418 60	0.836407	0.903846
P424	0.930193	0.884616
P427	0.849136	0.884616
P430	0.829077	0.942307
P503	0.894007	0.961538
P624917040852	0.898048	0.961538
P624917060027	0.936271	0.923077
P624917060157	0.825782	0.865385
P628747100045	0.810335	0.923077
P631257534507	0.876968	0.884616
P635824000013	0.801448	0.961538
P693749015017	0.886347	1

Table 4-1: A sample of time series clustering data sets

4.1.2 Time series clustering algorithms

Many time series clustering algorithms are available in inventory management. In this study, we applied two time series clustering algorithms: Expectation Maximization (EM) algorithm and K-Means.

K-Means is a commonly used time series clustering algorithm, described in section 2.1.3.1. It is a simple and fast clustering algorithm that attempts to minimize the sum of Euclidean distance between data objects in a cluster and the cluster center (Bradley et al., 1998; Bradley et al., 1998). It assumes that every data object belongs to exactly one cluster. Thus, no cluster overlap is allowed. K-Means algorithm requires the specification of the number of output clusters. Since the number of output clusters defines the number of groups that are produced by clustering tasks, it could be a challenge to properly define the number of output clusters.

EM algorithm is another popularly used time series clustering technique (Bradley et al., 1998; Bradley et al., 1998). It is a well-known method for estimating mixture model parameters, such as cluster parameters and their mixture weights. EM does not require the specification of distance measures and the number of output clusters. It iteratively refines model parameters, including the number of output clusters and terminates at a locally optimal solution. However, the locally optimal solution together with the number of output clusters may not be descriptive from a managerial point of view. This could be caused by the mismatch between statistical clustering results and management goals. For example, an optimal clustering solution may indicate that all data objects are included in one big group. It is statistically correct, but does not help the management goals. In such a situation, managerial adjustments are required to define a reasonable number of output clusters. Compared with other clustering algorithms, such as K-Means, the amount of computation involved with EM is relatively high.

4.1.3 Level-1 stability analysis – weekly analysis

The Level-1 stability analysis is also called weekly stability analysis. If products' sales quantities are approximately the same in all weeks, they are called stable weekly products. In this section, we perform two tiers of clustering analysis to identify weekly stable products.

4.1.3.1 Tier-1 weekly stability analysis

The goal of the Tier-1 time series clustering is to define stable groups from the noisy data set. The stability analysis starts with the EM algorithm. The analysis results are discussed below.

Table 4-2 shows the EM clustering results of the Tier-1 2005 weekly stability analysis. According to the results, 5737 products are clustered based on their weekly sales quantities and 4 groups (clusters) are generated as an optimal clustering solution. Among them, Cluster 3 has the highest value of P (0.3643), which means that products in this group were sold more often than products in other groups. Since the P values of Clusters 0, 1 and 2 are low, as 0.125, 0.0263 and 0.0665, respectively, products in these groups are not frequently sold products. Thus, Cluster 3 is the most stable group. That is, 1417 products, which are 25% of total products, are considered stable products in the Tier-1weekly stability analysis. The Tier-2 clustering analysis will be performed on these products.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	1203(21%)	0.6153	0.125
Cluster 1	2067(36%)	0.3105	0.0263
Cluster 2	1050(18%)	0.4796	0.0665
Cluster 3	**1417(25%)**	**0.7785**	**0.3643**

Table 4-2: Tier-1 weekly stability analysis with the EM algorithm in 2005

The EM clustering results of the Tier-1 2006 weekly stability analysis is shown in Table 4-3. We can see that 7525 products are clustered based on their weekly sales quantities and 5 groups (clusters) are generated as an optimal clustering solution. Similar to the distribution of weekly stability analysis with the EM algorithm in 2005, Cluster 2 has the highest value of P (0.6053) and the P values of the other clusters are relatively low. Products in Cluster 2 are considered stable products. Therefore, 1237 products, which are 16% of total products, are carried onto the Tier-2 clustering analysis in 2006.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	1334(18%)	0.7951	0.2772
Cluster 1	1138(15%)	0.4752	0.066
Cluster 2	**1237(16%)**	**0.7749**	**0.6053**
Cluster 3	2316(31%)	0.3121	0.0264
Cluster 4	1500(20%)	0.6207	0.1302

Table 4-3: Tier-1 weekly stability analysis with the EM algorithm in 2006

4.1.3.2 Tier-2 weekly stability analysis

In the Tier-2 clustering analysis, time series clustering techniques, such as EM and K-Means, are applied. We start clustering normalized data sets with the EM algorithm. However, we use the K-Means algorithm as an alternative solution to facilitate the Tier-2 clustering tasks when EM clustering results are not descriptive. Moreover, the number of output clusters is defined as 5 to distinguish products into finer groups, such as very stable, relatively stable, normal, relatively unstable and unstable groups.

Table 4-4 shows the Tier-2 clustering results based on Cluster 3 (1417 products) in the Tier-1 2005 weekly stability analysis. The results show that Cluster 11 is the most stable group since it has the highest P value (0.9128) and reasonable A value (0.7878). Therefore, 32 products are categorized as weekly stable products in 2005.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	203(14%)	0.8553	0.3763
Cluster 1	53(4%)	0.7295	0.3506
Cluster 2	129(9%)	0.7542	0.5721
Cluster 3	35(2%)	0.5943	0.4815
Cluster 4	86(6%)	0.7501	0.1779
Cluster 5	55(4%)	0.7815	0.297
Cluster 6	29(2%)	0.847	0.6301
Cluster 7	53(4%)	0.8544	0.4777
Cluster 8	101(7%)	0.7243	0.1974
Cluster 9	99(7%)	0.7491	0.225
Cluster 10	140(10%)	0.8369	0.2787
Cluster 11	**32(2%)**	**0.7878**	**0.9128**
Cluster 12	65(5%)	0.813	0.7518
Cluster 13	80(6%)	0.8019	0.5042
Cluster 14	53(4%)	0.6704	0.2531
Cluster 15	83(6%)	0.7795	0.205
Cluster 16	121(9%)	0.8061	0.2291

Table 4-4: Tier-2 weekly stability analysis with the EM algorithm in 2005

The Tier-2 clustering results of the Cluster 2 (1237 products) in the Tier-1 2006 weekly stability analysis are illustrated in Table 4-5. Comparing all the clusters, Cluster 9 represents the most stable group since the values of *A* (0.6061) and *P* (0.8791) are outstanding. Therefore, 21 products are categorized as weekly stable products in 2006.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	21(2%)	0.695	0.4107
Cluster 1	201(16%)	0.8263	0.4856
Cluster 2	140(11%)	0.837	0.4265
Cluster 3	127(10%)	0.8032	0.7054
Cluster 4	164(13%)	0.7944	0.8302
Cluster 5	180(15%)	0.7126	0.6467
Cluster 6	135(11%)	0.7762	0.9572
Cluster 7	197(16%)	0.7968	0.5789
Cluster 8	51(4%)	0.57	0.4105
Cluster 9	**21(2%)**	**0.6061**	**0.8791**

Table 4-5: Tier-2 weekly stability analysis with the EM algorithm in 2006

4.1.4 Level-2 stability analysis – monthly analysis

Products are considered stable monthly if their sales quantities are approximately the same in all months. In the Level-1 stability analysis, stable weekly products and unstable weekly products are distinguished from original data sets. Unstable weekly products are analyzed in the Level-2 stability analysis, also named monthly stability analysis. Moreover, stable weekly products can be also considered stable monthly products. Similar data mining tasks in weekly stability analysis are performed in monthly stability analysis. That is, we perform two-tier clustering analyses with EM and K-Means algorithms.

4.1.4.1 Tier-1 monthly stability analysis

Table 4-6 shows the EM clustering results of the Tier-1 2005 monthly stability analysis. In total, 5705 products are clustered based on their monthly sales quantities. The optimal clustering solution categorizes products into 7 groups (clusters). Cluster 0 has the highest value of P (0.9301), which means products in this group were sold more often than the others. The A value (0.8201) in Cluster 0 is reasonable. Thus, products in Cluster 0 are the most stable products in terms of monthly sales quantities in 2005. Therefore, 644 products, which are 11% of total products, are to be analyzed in the Tier-2 monthly stability analysis.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	**644(11%)**	**0.8201**	**0.9301**
Cluster 1	845(15%)	0.8406	0.2871
Cluster 2	1543(27%)	0.5528	0.0833
Cluster 3	878(15%)	0.7252	0.1765
Cluster 4	262(5%)	0.6947	0.6682
Cluster 5	719(13%)	0.9273	0.4137
Cluster 6	814(14%)	0.8359	0.6142

Table 4-6: Tier-1 monthly stability analysis with the EM algorithm in 2005

Table 4-7 shows the EM clustering results of the Tier-1 2006 monthly stability analysis. The results show that 7504 products are clustered based on their monthly sales quantities and 11 groups (clusters) are categorized through EM clustering algorithm. Cluster 5 has the highest value of P (0.9832) followed by Cluster 2 (0.973) and Cluster 10 (0.9679). Then, we compared A values of these Cluster 2, Cluster 5 and Cluster 10. The A value of Cluster 5 (0.7161) is the lowest one, so products in Cluster 5 are the most stable monthly products in 2006. Therefore, 237 products, which are 3% of total products, are to be analyzed in the Tier-2 monthly stability analysis.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	1203(21%)	0.7075	0.7118
Cluster 1	2067(36%)	0.8426	0.2734
Cluster 2	1050(18%)	0.8639	0.973
Cluster 3	1417(25%)	0.5528	0.0833
Cluster 4	679(9%)	0.8434	0.6143
Cluster 5	**237(3%)**	**0.7161**	**0.9832**
Cluster 6	539(7%)	0.8373	0.792
Cluster 7	815(11%)	0.9207	0.4173
Cluster 8	319(4%)	0.7613	0.4076
Cluster 9	925(12%)	0.7238	0.1724
Cluster 10	511(7%)	0.7951	0.9679

Table 4-7: Tier-1 monthly stability analysis with the EM algorithm in 2006

4.1.4.2 Tier-2 monthly stability analysis

The Tier-2 EM clustering results of the Cluster 1 (644 products) in the Tier-1 2005 monthly stability analysis is illustrated in Table 4-8. Statistically, the results show that one big group of all the products is the optimal solution. It doesn't meet the goal of the Tier-2 stability analysis, which refines groups. In this case, we use the K-Means algorithm to categorize products into 5 groups. The K-Means clustering results are shown in Table 4-9. According to the results, Cluster 2 is the most stable group since it's

A value (0.7664) is the lowest. Therefore, 87 products are categorized as stable monthly products in 2005.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	644(100%)	0.8183	0.9301

Table 4-8: Tier-2 monthly stability analysis with the EM algorithm in 2005

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	150(23%)	0.7802	1
Cluster 1	109(17%)	0.8483	0.9167
Cluster 2	**87(14%)**	**0.7664**	**0.9167**
Cluster 3	172(27%)	0.821	0.8333
Cluster 4	126(20%)	0.87	1

Table 4-9: Tier-2 monthly stability analysis with the K-Means algorithm in 2005

The Tier-2 EM clustering results of the Cluster 5 (237 products) in the Tier-1 2006 monthly stability analysis are shown in Table 4-10. The results show that Cluster 11 is the most stable group since its *A* value (0.5977) is the lowest. However, the number of products (8) in Cluster 11 is very small. In addition, many other groups, such as Clusters 7, 8, 9, etc, which also have reasonably good values of *A*, may contain stable monthly products. The EM clustering results are not descriptive as the boundaries of groups are not clear. Thus, we apply the K-Means algorithm to categorize these products into 5 groups so that we can distinguish groups based on their stability levels. The K-Means clustering results of these 237 products are shown in Table 4-11. Cluster 3 provides a reasonable group of stable monthly products. Therefore, 26 products are categorized as monthly stable products in 2006.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	18(8%)	0.7063	1
Cluster 1	22(9%)	0.6941	1
Cluster 2	13(5%)	0.7074	0.9167
Cluster 3	32(14%)	0.7328	1
Cluster 4	26(11%)	0.7183	1
Cluster 5	4(2%)	0.6819	0.9167
Cluster 6	17(7%)	0.7272	0.9167
Cluster 7	9(4%)	0.6578	0.9167
Cluster 8	17(7%)	0.6757	1
Cluster 9	16(7%)	0.6505	1
Cluster 10	46(19%)	0.7425	1
Cluster 11	**8(3%)**	**0.5977**	**1**
Cluster 12	9(4%)	0.7496	1

Table 4-10: Tier-2 monthly stability analysis with the EM algorithm in 2006

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	102(43%)	0.7361	1
Cluster 1	13(5%)	0.6642	0.9167
Cluster 2	66(28%)	0.6968	1
Cluster 3	**26(11%)**	**0.6384**	**1**
Cluster 4	30(13%)	0.7186	0.9167

Table 4-11: Tier-2 monthly stability analysis with the K-Means algorithm in 2006

4.1.5 Level-3 stability analysis – quarterly analysis

Products, which have stable sales quantity in each quarter, are considered quarterly stable products. In addition, stable weekly and stable monthly products can be considered quarterly stable products. Here, the Level-3 stability analysis, which is also known as quarterly stability analysis, is applied on unstable monthly products. Again, two tiers of analyses with EM and K-Means algorithms are performed in the Tier-2 quarterly stability analysis.

4.1.5.1 Tier-1 quarterly stability analysis

Tables 4-12 and 4-13 illustrate EM clustering results of the Tier-1 quarterly stability analysis in 2005 and 2006. The optimal clustering results show that only two clusters (groups) are distinguished from data sets in 2005 and 2006. Moreover, the stable groups have 2689 (48%) and 4150 (55%) products in 2005 and 2006, respectively. Statistically, the results are correct as clustering data objects based on their statistic values. However, they do not meet the reality of the real world inventory management. In the real world, stable products do not represent high proportions of total products. This could be caused by the fact that the value of P is limited as 0, 0.25, 0.5, 0.75 and 1.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	**2689(48%)**	**0.8472**	**0.898**
Cluster 1	2929(52%)	0.8944	0.354

Table 4-12: Tier-1 quarterly stability analysis with the EM algorithm in 2005

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	**4150(55%)**	**0.856**	**0.9211**
Cluster 1	3328(45%)	0.8924	0.3537

Table 4-13: Tier-1 quarterly stability analysis with the EM algorithm in 2006

As an alternative time-series clustering algorithm, K-Means is applied on quarterly stability analysis in 2005 and 2006. We categorize products into 7 groups (clusters). In this way, we could distinguish products based on their stability levels, such as very stable, relatively stable, stable, normal, unstable, relatively unstable and very unstable. Tables 4-14 and 4-15 show K-Means clustering results of quarterly stability analysis in 2005 and 2006. According to the results, in 2005, Cluster 2 is a very stable quarterly group since it has the optimal P value (1) and an acceptable A value (0.8102). Cluster 5 can be considered another very stable group since it's A value is the lowest and P value (0.7764) is acceptable. To avoid losing stable quarterly products, 777+218=995 products will be analyzed in the Tier-2 2005 quarterly stability analysis. In 2006, Cluster 5 represents an excellent group, which has the optimal values of A (0) and P (1). Here, the P value (1) means that products are sold in every quarter and the A value (0) means that products are sold exactly the same number of times in each quarter in 2006. Therefore, the clustering results for quarterly stability analysis in 2006 is finalized, no further clustering is required. 32 products are categorized as stable quarterly products in 2006.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	563(10%)	0.8331	0.75
Cluster 1	1736(31%)	0.866	0.25
Cluster 2	**777(14%)**	**0.8102**	**1**
Cluster 3	1193(21%)	0.9364	0.5
Cluster 4	799(14%)	0.9207	1
Cluster 5	**218(4%)**	**0.6325**	**0.7764**
Cluster 6	332(6%)	0.9201	0.75

Table 4-14: Tier-1 quarterly stability analysis with the K-Means algorithm in 2005

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	680(9%)	0.9692	1
Cluster 1	1389(19%)	0.8693	1
Cluster 2	1986(27%)	0.866	0.25
Cluster 3	1319(18%)	0.8435	0.75
Cluster 4	730(10%)	0.7855	1
Cluster 5	**32(0%)**	**0**	**1**
Cluster 6	1342(18%)	0.9319	0.5

Table 4-15: Tier-1 quarterly stability analysis with the K-Means algorithm in 2006

4.1.5.2 Tier-2 quarterly stability analysis

The Tier-2 EM clustering results of these 995 products in the Tier-1 2005 quarterly stability analysis are shown in Table 4-16. Obviously, Cluster 1 is the most stable group because it has the optimal values of A (0) and P (1). Therefore, 23 products are categorized as stable quarterly products in 2005.

Product groups	Number of products (percentage)	Mean of absolute z-score (A)	Percentage of non-zero values (P)
Cluster 0	195(20%)	0.7071	0.75
Cluster 1	**23(2%)**	**0**	**1**
Cluster 2	0(0%)	0.8527	1
Cluster 3	691(69%)	0.823	1
Cluster 4	86(9%)	0.7071	0.9339

Table 4-16: Tier-2 quarterly stability analysis with the EM algorithm in 2005

As a result, 6 stable groups are obtained from stability analysis in 2005 and 2006. Table 4-17 lists the stable groups. Their stability levels and the number of products are also included in the table. In total, we have 142 stable products in 2005 and 79 stable products in 2006. In addition, we also compared stable products in 2005 with stable products in 2006 at three levels. There are 3 stable weekly products and 3 stable monthly products sold in both 2005 and 2006. No stable quarterly products were sold in both the years.

Stable group	Stable period	Number of products
2005-1	Weekly	32
2005-2	Monthly	87
2005-3	Quarterly	23
Total 2005		**142**
2006-1	Weekly	21
2006-2	Monthly	26
2006-3	Quarterly	32
Total 2006		**79**

Table 4-17: List of stable groups in 2005 and 2006

4.2 Seasonality analysis

Products are considered seasonal products if their sales quantities periodically change with time. As we discussed in section 3.1.3, if products' selling ratios in one period are high, they are called single period selling products. In addition, products, which were mostly sold in two periods, are called double-periods selling products. Based on products' sales patterns, we propose a couple of inventory management strategies to control seasonal inventories.

(i) Carry very few seasonal products in their off-sales periods. The inventory in off-season should be based on quantities from previous year sales during off-season.

(ii) Order a lot of seasonal products in their sales periods. . The size of order should be based on quantities from previous year sales during the same season.

Similar normalization processes, described in section 4.1.1, are applied on monthly and quarterly sales quantities data. Here, we normalize data sets using two methods: max-normalization that divides the maximum object value and mean-normalization that divides the mean object value. In this section, we perform two levels of seasonality analyses, including monthly and quarterly seasonality analysis, to discover products' sales patterns. Moreover, two tiers of clustering analysis are performed for each level of seasonality analyses. That is, the Tier-1 clustering analysis indicates reasonable seasonal groups and the Tier-2 clustering analysis categorizes these seasonal groups into finer groups. Since the amount of computation in seasonality analysis is huge, we choose the K-Means algorithm to facilitate seasonality analysis.

In addition, some of the products may belong to multiple clusters to a certain extent. Such situations can be handled using soft clustering techniques, such as fuzzy and rough clustering. However, we will restrict our studies to the traditional crisp clustering algorithm.

4.2.1 Level-1 seasonality analysis – monthly analysis

The Level-1 seasonality analysis, which is also named monthly seasonality analysis, is performed based on products' monthly sales quantity. We discover products' monthly selling patterns in this section. Products are considered single month selling products if they were mostly sold in one month. Similarly, products that were mostly sold in two months are considered two-months selling products.

4.2.1.1 Tier-1 monthly seasonality analysis

In the Tier-1 monthly seasonality analysis, some possible groups are 6 single month selling groups, 3 double sales-month groups and 1 random selling group. Thus, we categorize products into 10 reasonable groups. Compared to clustering results based on max-normalization, seasonality analysis associated with mean-normalization method provides more distinct results. Some of these clustering results are discussed in this section.

Tables 4-18 shows the Tier-1 seasonality analysis results based on monthly sales quantities in 2005. Cluster 1 is a typical single month selling group in December since its December value (11.6637) is extremely high and the rest of the values are inane. The proposed inventory management strategies can be applied on single month selling products. Based on the historical product demand, the store can (i) carry a tiny amount of these products in off-selling months, which are from January to November, and (ii) store an adequate amount of these inventories in the selling month, which is December. Similarly, Clusters 0, 2, 3, 4 and 9 are single month selling groups in October, February, January, August and June, respectively. In addition, Cluster 5 can be considered as a typical example of a three-months selling group since the sales quantities in July, September and November are dominating. Products in these seasonal groups can be controlled with the proposed inventory management strategies. Moreover, we express 2005 monthly sales trends in Figures 4-1. In such a way, we can identify seasonal selling groups clearly and confirm our findings above.

Product groups	Number of products (percentage)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cluster 0	137(2%)	-0.0584	0	0	0.0876	0	0.0584	0.1168	0.2628	0.0438	**11.1888**	0.2088	0.0915
Cluster 1	333(6%)	0.0742	0.0921	-0.0284	0.0177	0.0098	0.0019	-0.0069	0.0781	-0.009	0.0739	0.0328	**11.6637**
Cluster 2	263(5%)	0.603	**9.684**	0.3938	0.4148	0.371	0.0423	0.1451	0.0754	0.0306	0.0371	0.132	0.071
Cluster 3	216(4%)	**10.0978**	0.3158	0.0893	0.5497	0.259	0.0417	0.2913	-0.0195	0.07	0.0157	0.2234	0.0657
Cluster 4	178(3%)	0.2783	0.2801	0.3844	0.2607	0.3567	0.0417	0.4221	**9.1954**	0.1335	0.0892	0.4068	0.1511
Cluster 5	647(12%)	0.2486	0.2233	0.1011	0.3011	0.3212	0.1453	**2.8481**	0.3915	**3.8863**	0.2118	**3.1742**	0.1474
Cluster 6	690(12%)	0.756	0.603	0.6205	0.631	0.7603	0.5069	0.8187	0.781	0.7852	0.3863	1.1929	4.1582
Cluster 7	642(11%)	0.5821	0.515	0.5615	0.6527	0.7304	0.4394	0.7624	0.8104	0.8001	3.7798	1.2038	1.1623
Cluster 8	2191(39%)	1.0103	0.9072	1.8872	1.5889	1.6778	0.7572	0.9695	0.8426	0.5937	0.4894	0.7004	0.5757
Cluster 9	298(5%)	0.4895	0.4261	0.3657	0.5268	0.7044	**7.0967**	0.5732	0.4903	0.298	0.1779	0.4095	0.4419

Table 4-18: Tier-1 monthly seasonality analysis in 2005

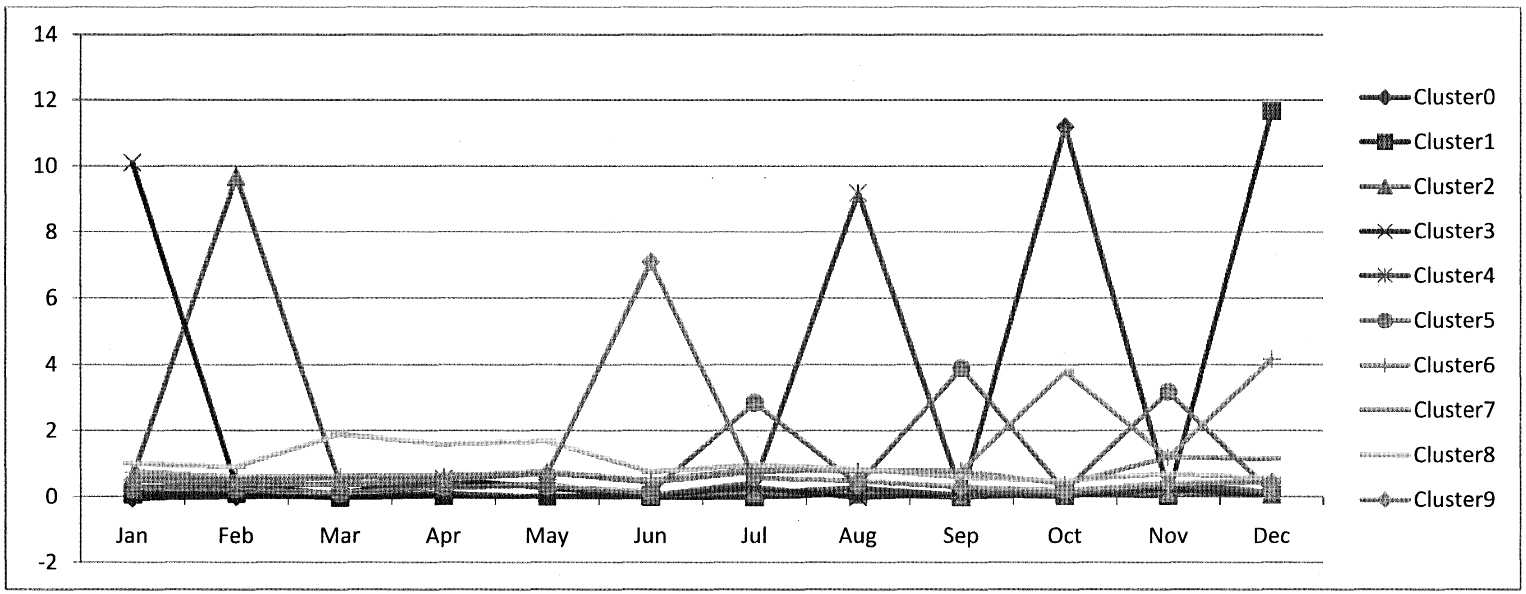


Figure 4-1: Tier-1 monthly seasonality analysis - 2005 monthly sales trend

Similarly, the Tier-1 2006 monthly seasonality analysis results are illustrated in Table 4-19. In 2006, Cluster 5 is a single month selling group in January since the normalized sales quantity in January (11.4969) is extremely high. The numbers of products sold in their selling month (January) are extremely high. In addition, these products were rarely sold in off-selling months. Thus, the proposed inventory management strategies can be applied on these products. According to historical sales records, the store can (i) keep a very small amount of these products in off-selling months and (ii) order sufficient inventories in the selling month, which is January in this case. Clusters 0, 1, 3, 6 and 9 are also single month selling products in May, June, November, July and March, 2006. Products in Cluster 7 were mostly sold in two-months: April (4.581) and August (5.0619). The inventory management strategies can also be applied to these two-months selling products. That is, double-month selling products should be kept at a reasonable level for two-months instead of one month and the number of inventories should refer to sales quantities in both of their selling months respectively. Cluster 2 seems to have a high sales quantity in December. In addition, quite a few sales were made in February. This could be a typical fuzzy group. The Tier-2 monthly seasonality analysis may categorize these products into finer groups. Figure 4-2 illustrates sales trends of each group in the Tier-1 2006 monthly seasonality analysis. It visually confirms our findings and supports the proposed inventory management strategies.

Product groups	Number of products (percentage)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cluster0	216(3%)	0.4886	0.3621	0.1009	0.8347	**8.1574**	0.0424	0.3796	0.3955	0.3898	0.2057	0.2845	0.3587
Cluster1	300(4%)	0.334	0.2285	0.4316	0.2531	0.3334	**8.2209**	0.3823	0.2791	0.4339	0.2419	0.3334	0.5278
Cluster2	514(7%)	0.3665	2.4443	0.092	0.2049	0.0791	0.0958	0.0682	0.1879	0.2918	0.4224	0.3858	7.3612
Cluster3	341(5%)	0.4291	0.1782	0.0126	0.2092	0.0745	0.0701	0.093	0.4108	0.3262	0.6393	**8.5384**	1.0185
Cluster4	3599(48%)	1.512	1.0155	1.0241	0.9579	0.9209	0.8939	0.5325	0.7946	0.8718	1.4709	0.9978	1.008
Cluster5	564(8%)	**11.4969**	0.0777	0.0369	0.0502	0.0209	0.034	0.0679	0.0242	0.0594	0.0064	0.0525	0.0729
Cluster6	114(2%)	0.0886	0.0026	0.1003	0.0566	0.1171	0.1943	**10.9792**	0.148	0.0872	0.0406	0.0763	0.1092
Cluster7	354(5%)	0.2912	0.2895	0.1506	**4.8473**	0.0779	0.1621	0.0879	**5.1765**	0.4115	0.2417	0.1457	0.1181
Cluster8	1157(16%)	0.7566	0.4381	0.4295	0.6801	0.6475	0.8398	2.5253	1.0027	2.3278	0.7339	0.803	0.8157
Cluster9	287(4%)	0.6418	0.7014	**8.1342**	0.5225	0.484	0.124	0.2118	0.3297	0.3309	0.1765	0.2025	0.1408

Table 4-19: Tier-1 monthly seasonality analysis in 2006

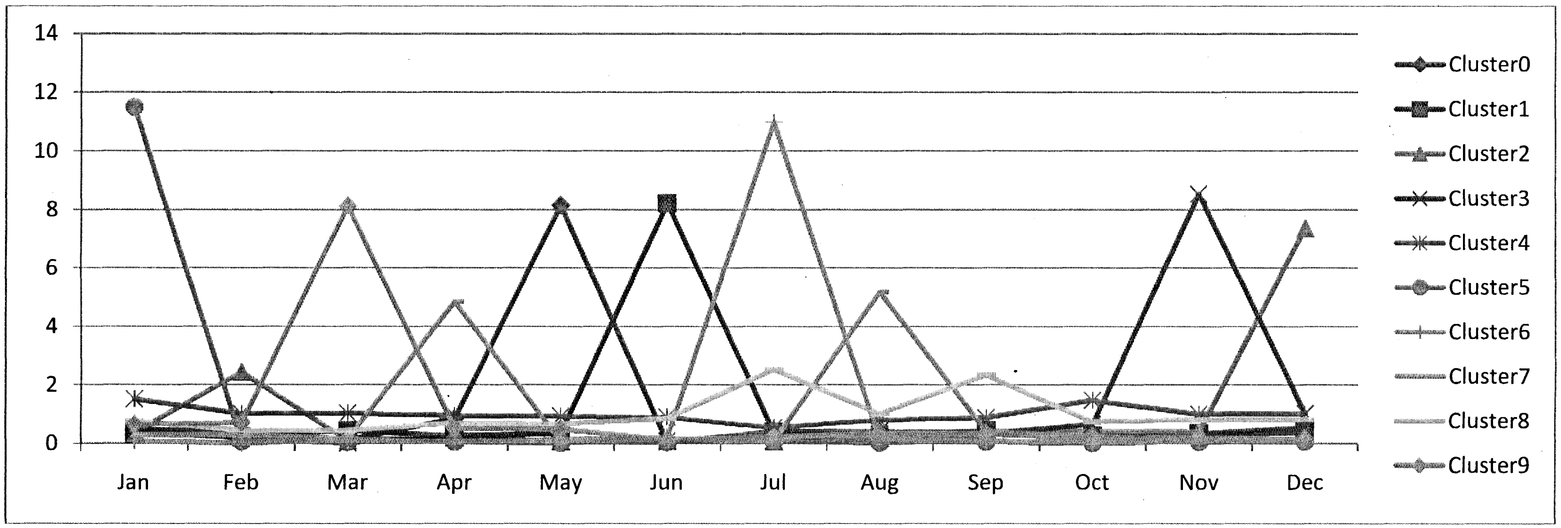


Figure 4-2: Tier-1 monthly seasonality analysis - 2006 monthly sales trend

4.2.1.2 Tier-2 monthly seasonality analysis

The Tier-1 clustering analysis above provides us with several reasonable groups. The Tier-2 clustering analysis is performed based on these groups so that we can identify their sales patterns and categorize them into finer groups. Here, we categorize products into 5 groups based on their seasonalities in monthly sales quantities. A typical Tier-2 monthly seasonality analysis results are discussed below.

The Tier-1 2006 monthly seasonality analysis indicates that Cluster 2 is a fuzzy group. Products in Cluster 2 are further analyzed in the Tier-2 monthly seasonality analysis and the clustering results are shown in Table 4-20. The results show that products in Cluster 1 are single month selling products in February, 2006. The proposed seasonal inventory management strategies can be applied on these single month selling products. That is, the store can keep a high amount of these products in February and carry only a few of them for the rest of year. Cluster 0 is also a single month selling group, where products were sold mostly in December. Clusters 2, 3 and 4 are really two-months selling groups in October-December, January-December and August-December, respectively. Products in these groups can be managed with our seasonal inventory management strategies. All 5 clusters are plotted in Figure 4-3 so that we can easily identify products' sales patterns through the graph. The figure confirms our findings above and supports the proposed inventory management strategies.

Product groups	Number of products (percentage)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cluster 0	328(64%)	0.101	0.4512	0.1107	0.2901	0.0983	0.1284	0.0855	0.0519	0.3569	0.0734	0.4707	**9.782**
Cluster 1	92(18%)	0	**11.902**	0	0	0	0	0	0	0	0	0.0326	0.0652
Cluster 2	43(8%)	0.0856	0.0558	0.186	0	0.0558	0	0.1395	0	0.2601	**4.3521**	0.6004	**6.2645**
Cluster 3	31(6%)	**4.8906**	0.129	0.0968	0.0968	0	0	0	0.043	0.4731	0	0.1742	**6.0965**
Cluster 4	20(4%)	0	0.35	0	0.36	0.3	0.3557	0.05	**3.9105**	0.3545	0.2945	0.4848	**5.5399**

Table 4-20: Tier-2 2006 monthly seasonality analysis – Cluster 2

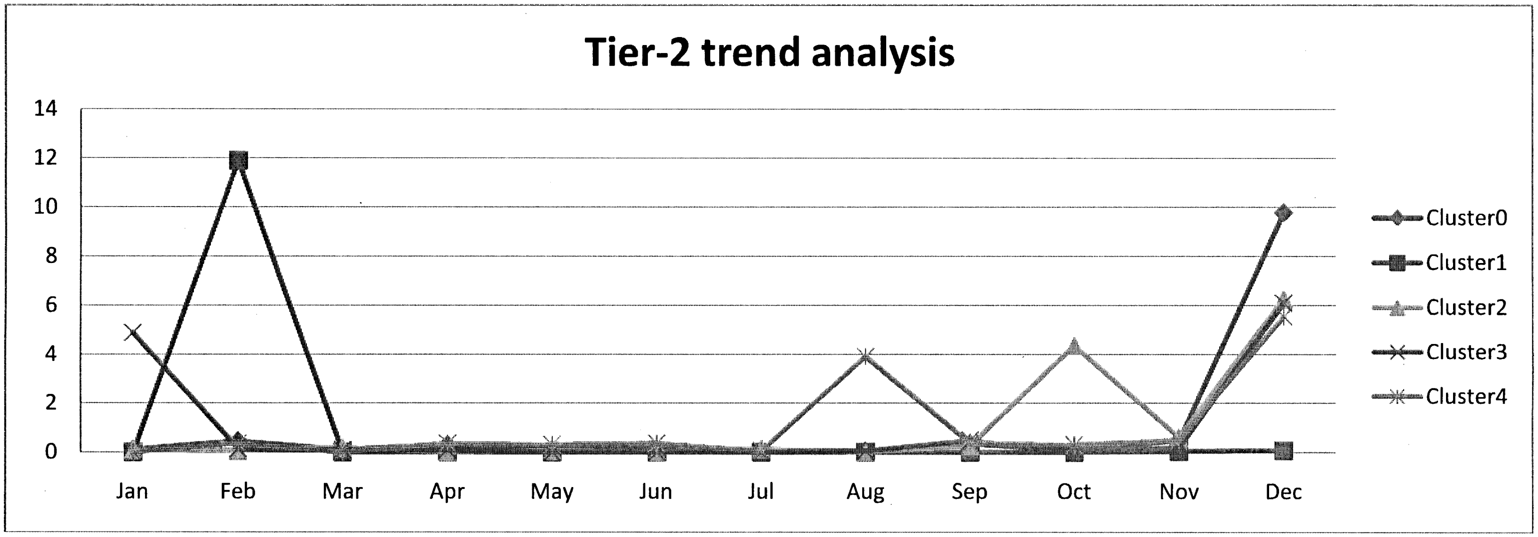


Figure 4-3: Tier-2 monthly seasonality analysis – 2006 monthly sales trend of Cluster 2

The monthly seasonality grouping results in 2005 and 2006 are shown in Tables 4-21 and 4-22. These Tables also describes processes to locate seasonal groups, sales months of seasonal groups and number of products included in the groups. In addition, sales months of two-months selling groups are listed based on priority. That is, higher sales months are listed first in the column. In total, there are 27 seasonal groups defined from the 2005 monthly seasonality analysis. These capture 1485 products. In 2006, 2194 products are categorized into 29 seasonal groups according to monthly seasonality analysis. The rest of products are analyzed in quarterly seasonality analysis.

Seasonal	Tier-1 cluster	Tier-2 cluster	Selling month(s)	Number of
1	0	0	October	106
2	0	1	October and November	23
3	0	2	October and July	4
4	0	3	October and August	2
5	0	4	October and June	2
6	1	0	December	298
7	1	1	December and February	5
8	1	2	December and November	4
9	1	4	December and February	7
10	2	0	February and March	20
11	2	1	February	154
12	2	2	February and January	36
13	2	4	February and August	3
14	3	0	January	179
15	3	1	January and July	10
16	3	2	January and November	7
17	3	3	January and February	17
18	3	4	January and June	3
19	4	1	August and February	9
20	4	2	April and August	6
21	4	3	August	90
22	4	4	August and March	11
23	5	0	November and July	256
24	5	2	September	103
25	9	1	June	78
26	9	2	June and February	25
27	9	4	June and November	27
Total				**1485**

Table 4-21: 2005 seasonal groups based on monthly seasonality analysis

Seasonal	Tier-1 cluster	Tier-2 cluster	Selling month(s)	Number of
1	0	0	May	76
2	0	2	May and July	16
3	0	3	May and August	12
4	0	4	May and September	14
5	1	1	June and August	20
6	1	2	June	109
7	1	4	June and December	27
8	2	0	December	328
9	2	1	November	92
10	2	2	December and October	43
11	2	3	December and January	31
12	2	4	December and September	20
13	3	0	November and August	21
14	3	1	November and July	9
15	3	3	November and September	25
16	3	4	November	138
17	5	0	January	527
18	5	1	January and March	10
19	5	2	January and December	13
20	6	0	July	84
21	6	1	July and November	2
22	6	2	July and May	3
23	6	3	July and June	6
24	7	0	April	129
25	7	2	August	152
26	8	4	September	137
27	9	1	March	103
28	9	2	March and April	24
29	9	4	March and August	23
Total				**2194**

Table 4-22: 2006 seasonal groups based on monthly seasonality analysis

4.2.2 Level-2 seasonality analysis –quarterly analysis

The Level-2 seasonality analysis is performed based on products' quarterly sales quantities. Two tiers of seasonality analysis are applied to categorize products into reasonable groups. In this section, we aim on finding out single quarter selling products and two-quarters selling products.

4.2.2.1 Tier-1 quarterly seasonality analysis

Some possible groups are 4 single sales-season groups, 2 double sales-season groups and 1 random selling group. Thus, we cluster products into 7 reasonable groups in the Tier-1 analysis. The Tier-1 2005 quarterly seasonality analysis results are shown in Table 4-23. Obviously, Clusters 0, 1 and 2 are single quarter selling groups in Quarter 2, 4 and 1, respectively. Products in these groups have relatively high sales quantities in their sales quarters. The seasonal inventory management strategies can be applied to control quantities of these products. That is, the store can carry a few of these products in off-sales quarters and order a lot in the sales quarter. Cluster 5 is an interesting group. Since products' quarterly sale quantities decreased throughout 2006, this might be a group of products, which have relatively short business lives. Products in the rest of the groups seem to have insignificant seasonal patterns. They may be illustrated more clearly in the Tier-2 quarterly seasonality analysis.

Product groups	Number of products (percentage)	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 0	347(8%)	0.1823	**3.5651**	0.2009	0.0517
Cluster 1	282(7%)	0.1142	0.1781	0.3746	**3.3331**
Cluster 2	272(7%)	**3.3824**	0.2923	0.1525	0.1727
Cluster 3	764(19%)	0.7123	0.3566	1.0456	1.8855
Cluster 4	521(13%)	0.3823	0.8343	2.3656	0.4178
Cluster 5	823(20%)	**1.6339**	**1.1845**	**0.7976**	**0.384**
Cluster 6	1101(27%)	0.7447	1.2506	0.7907	1.214

Table 4-23: Tier-1 2005 quarterly seasonality analysis

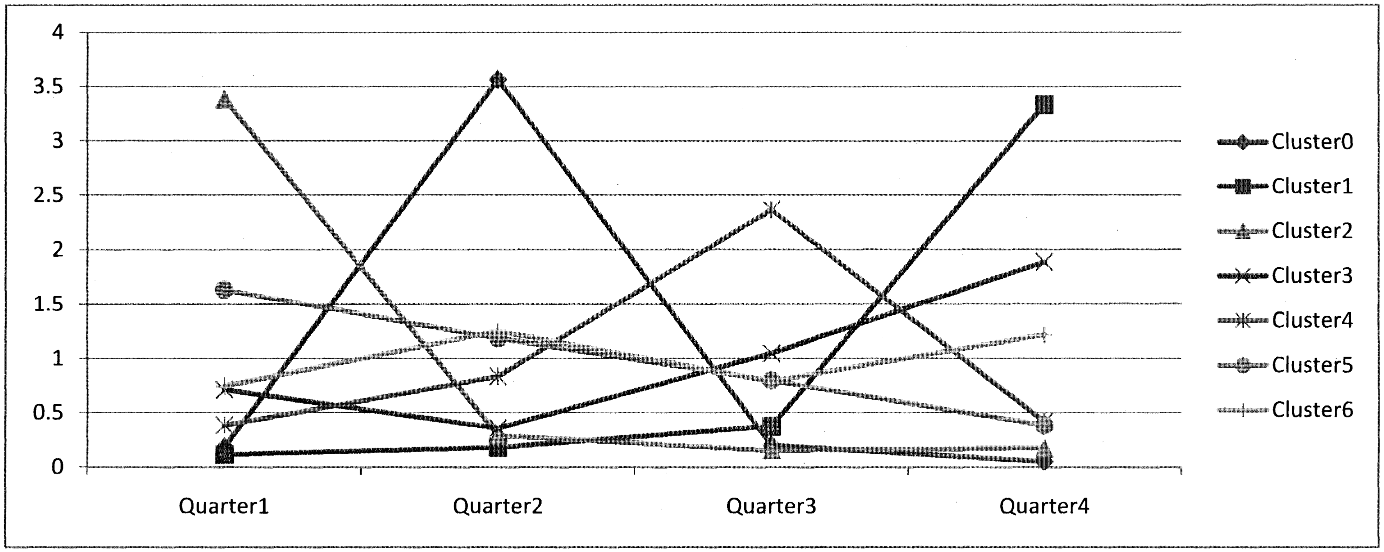


Figure 4-4: Tier-1 seasonality analysis – 2005 quarterly sales trend

Table 4-24 shows the Tier-1 2006 quarterly seasonality analysis results. According to the results, Cluster 2 has the most significant seasonal sales patterns. Products in this group were sold mostly in Quarter 4; they are single quarter selling products. The following groups are Clusters 4 and 0. They have clear sales patterns, but not as significant as Cluster 2. Cluster 6 is a tricky group. It can be considered single quarter selling group since products' sales quantities in the first quarter are relatively high. However, products' sales quantities in the first and second quarters are dominating in this group. That is, it is a two-quarters selling group in Quarters 1 and 2. In addition, products in this group were sold more in Quarter 1 than in Quarter 2. The Tier-2 quarterly seasonality analysis may identify products' sales patterns more clearly.

Product groups	Number of products (percentage)	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 0	434(8%)	0.3431	0.735	**2.4959**	0.4259
Cluster 1	1001(19%)	0.4543	0.7123	1.3548	1.4786
Cluster 2	387(7%)	0.255	0.1847	0.1689	**3.3914**
Cluster 3	981(19%)	1.6348	0.4959	0.9308	0.9385
Cluster 4	384(7%)	0.3653	**2.553**	0.3459	0.7358
Cluster 5	1560(30%)	1.0908	1.2294	0.7693	0.9106
Cluster 6	505(10%)	**2.7648**	**0.8854**	0.1177	0.2321

Table 4-24: Tier-1 2006 quarterly seasonality analysis

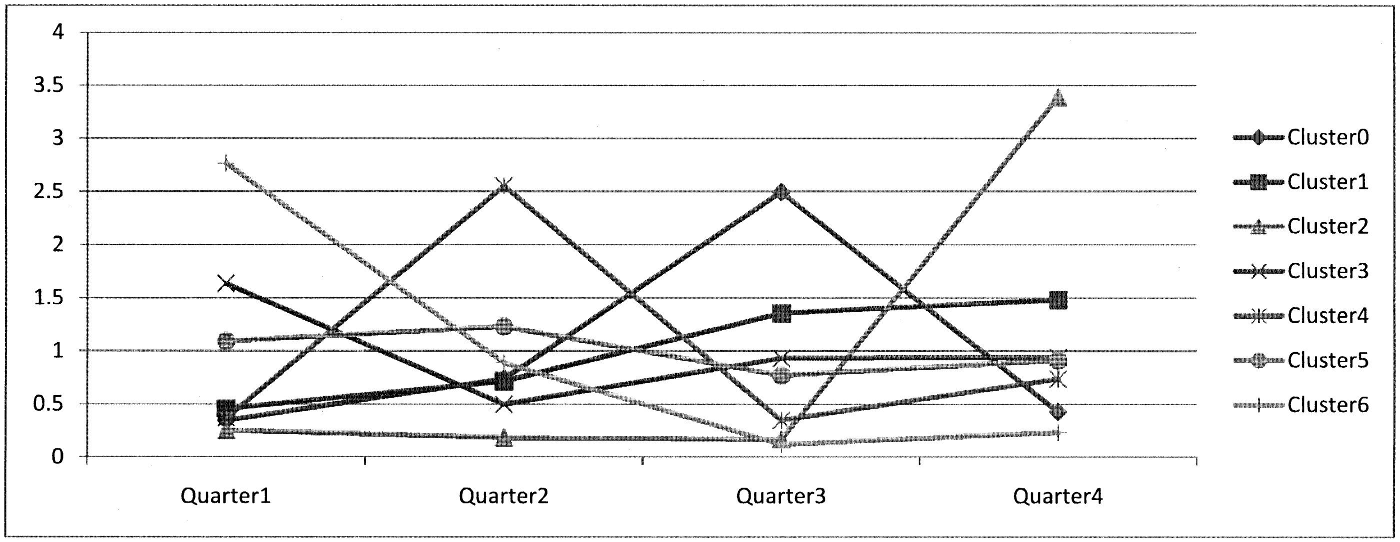


Figure 4-5: Tier-1 seasonality analysis – 2006 quarterly sales trend

4.2.2.2 Tier-2 quarterly seasonality analysis

Several reasonable groups are provided by the Tier-1 quarterly seasonality analysis. The Tier-2 clustering analysis refines the clustering results. Here, we cluster products into 5 groups based on their quarterly sales patterns. Some of these Tier-2 clustering results are discussed below.

According to the Tier-1 2006 quarterly seasonality analysis, Cluster 2 (387 products) is a single quarter selling group in Quarter 4. The Tier-2 quarterly seasonality analysis refines the grouping results as shown in Table 4-25. Obviously, Cluster 0 is a single quarter selling group with the most significant seasonal sales pattern. Products in this group were sold mostly in Quarter 4. Clusters 2 and 4 are two-quarters selling groups in Quarters 3-4 and Quarters 2-4, respectively. Coincidently, Clusters 1 and 3 are both two-quarters selling groups in Quarters 1-4. However, they are two seasonal groups because the weights of sales quantities between these two quarters are different. According to products' sales pattern in these seasonal groups, our seasonal inventory management strategies can be smoothly applied. Figure 4-6 illustrates the sales trends of these seasonal groups in Table 4-25. It graphically proves our findings above and supports the proposed inventory management strategies.

Product groups	Number of products (percentage)	Quarter1	Quarter2	Quarter3	Quarter4
Cluster 0	200(52%)	0.0083	0.0023	0.0137	**3.9757**
Cluster 1	24(6%)	**0.8134**	0.0181	0.0235	**3.145**
Cluster 2	70(18%)	0.2307	0.1275	**0.7696**	**2.8722**
Cluster 3	42(11%)	**1.1932**	0.1293	0.0969	**2.5806**
Cluster 4	51(13%)	0.2207	**1.1021**	0.0805	**2.5966**

Table 4-25: Tier-2 2006 quarterly seasonality analysis - Cluster 2

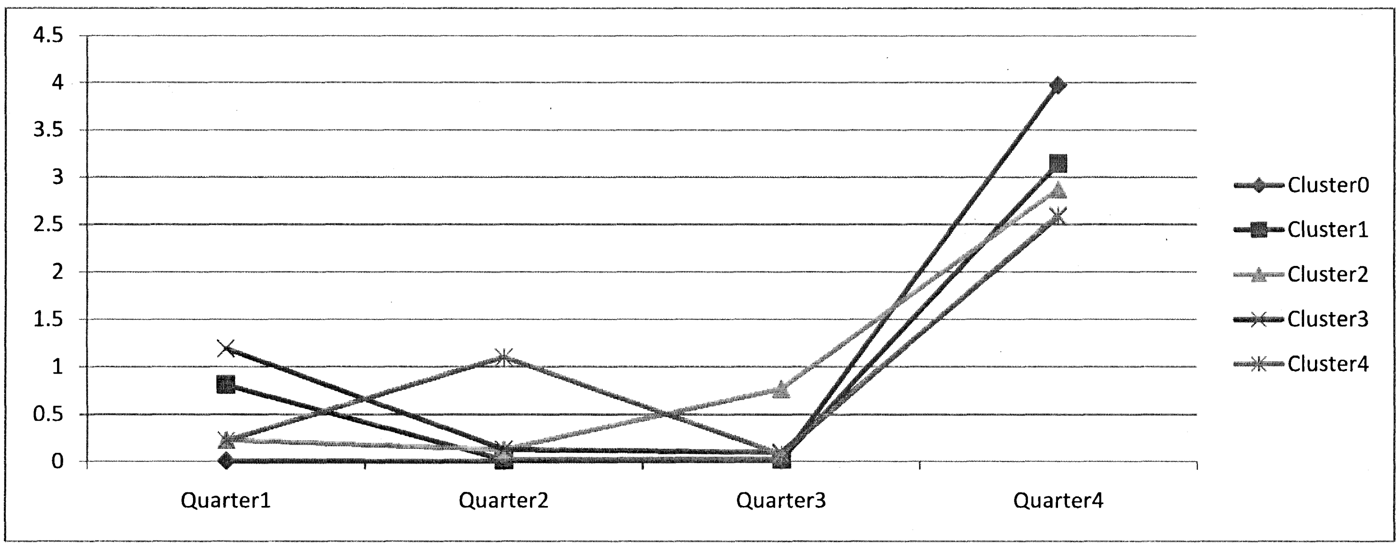


Figure 4-6: Tier-2 seasonality analysis – 2006 quarterly sales trend of Cluster 2

The consolidated quarterly seasonal grouping results in 2005 and 2006 are listed in Tables 4-26 and 4-27. Processes to identify seasonal groups, sales quarters of seasonal groups and number of products contained are also included in the Tables. Note, the primary sales quarter is shown firstly in two-quarters selling groups. In total, there are 23 quarterly seasonal groups identified in 2005. These include 1817 products. In 2006, 1770 products are categorized into 20 quarterly seasonal groups. The rest of products are considered random selling products.

Seasonal	Tier-1 cluster	Tier-2 cluster	Selling quarter(s)	Number of
28	0	0	Quarter 2	218
29	0	1	Quarters 2 and 3	45
30	0	2	Quarters 2 and 4	13
31	0	3	Quarters 2 and 1	55
32	0	4	Quarter 2	16
33	1	0	Quarters 4 and 3	46
34	1	1	Quarters 4 and 2	43
35	1	2	Quarters 4 and 3	51
36	1	3	Quarter 4	122
37	1	4	Quarters 4 and 1	20
38	2	0	Quarter 1	135
39	2	1	Quarters 1 and 4	18
40	2	2	Quarter 1	21
41	2	3	Quarters 1 and 4	12
42	2	4	Quarters 1 and 2	86
43	3	0	Quarters 4 and 3	213
44	3	2	Quarters 1 and 4	112
45	4	1	Quarters 3 and 4	72
46	4	3	Quarters 3 and 2	129
47	4	4	Quarter 3	54
48	5	3	Quarters 1 and 3	107
49	5	4	Quarters 1 and 2	149
50	6	1	Quarters 2 and 4	80
Total				**1817**

Table 4-26: 2005 seasonal groups based on quarterly seasonality analysis

Seasonal	Tier-1 cluster	Tier-2 cluster	Selling quarter(s)	Number of
30	0	1	Quarters 3 and 2	77
31	0	2	Quarters 3 and 1	98
32	0	3	Quarter 3	53
33	0	4	Quarters 3 and 4	76
34	1	1	Quarters 4 and 3	169
35	2	0	Quarter 4	200
36	2	1	Quarters 4 and 1	24
37	2	2	Quarters 4 and 3	70
38	2	3	Quarters 4 and 1	42
39	2	4	Quarters 4 and 2	51
40	3	2	Quarters 1 and 3	127
41	3	3	Quarters 1 and 4	111
42	4	0	Quarters 2 and 1	100
43	4	1	Quarters 2 and 4	98
44	4	2	Quarter 2	64
45	4	4	Quarters 2 and 3	44
46	6	0	Quarters 1 and 2	43
47	6	2	Quarters 1 and 2	120
48	6	3	Quarters 1 and 2	160
49	6	4	Quarter 1	43
Total				**1770**

Table 4-27: 2006 seasonal groups based on quarterly seasonality analysis

4.3 Summary of product analyses

In this study, we performed stability and seasonality analysis on a retail chain data set. The EM and K-Means clustering algorithms are used to facilitate product analyses. According to the analyses results, 5737 products were categorized into 54 groups in 2005. These included:

- 3 stable groups
    - 1 stable weekly group
    - 1 stable monthly group
    - 1 stable quarterly group
- 50 seasonal groups
    - 6 single-month selling groups
    - 21 two-month selling groups
    - 6 single-quarter selling groups
    - 17 two-quarters selling groups.
- 1 random group

In 2006, 7525 products were categorized into 53 groups. These included:

- 3 stable groups
    - 1 stable weekly group
    - 1 stable monthly group
    - 1 stable quarterly group
- 49 seasonal groups
    - 11 single-month selling groups
    - 18 two-month selling groups

- 4 single-quarter selling groups
- 16 two-quarters selling groups.

➢ 1 random group

Figures 4-7 and 4-8 illustrate product distributions in 2005 and 2006, respectively. Product groups and numbers of products they contained are labelled in the Figures. We can see that sales patterns of over 50% of products have been identified through product analyses. Moreover, profits distributions in 2005 and 2006 are described in Figures 4-9 and 4-10. We can see that stable and seasonal products made significant contributions to the store's profits in 2005 and 2006.

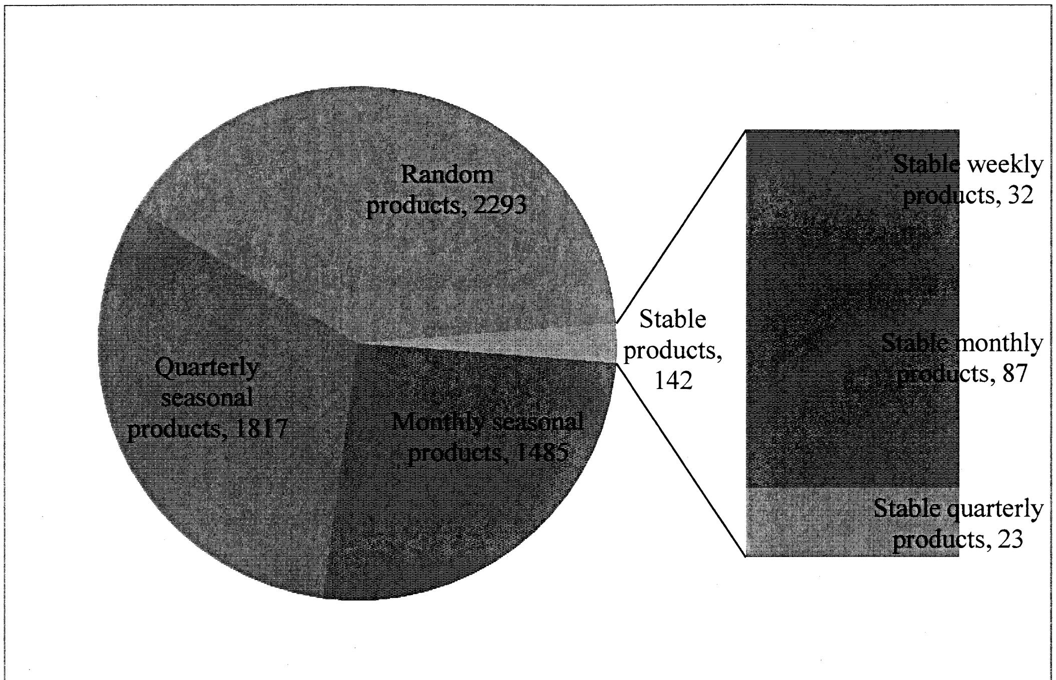


Figure 4-7: Product distribution in 2005

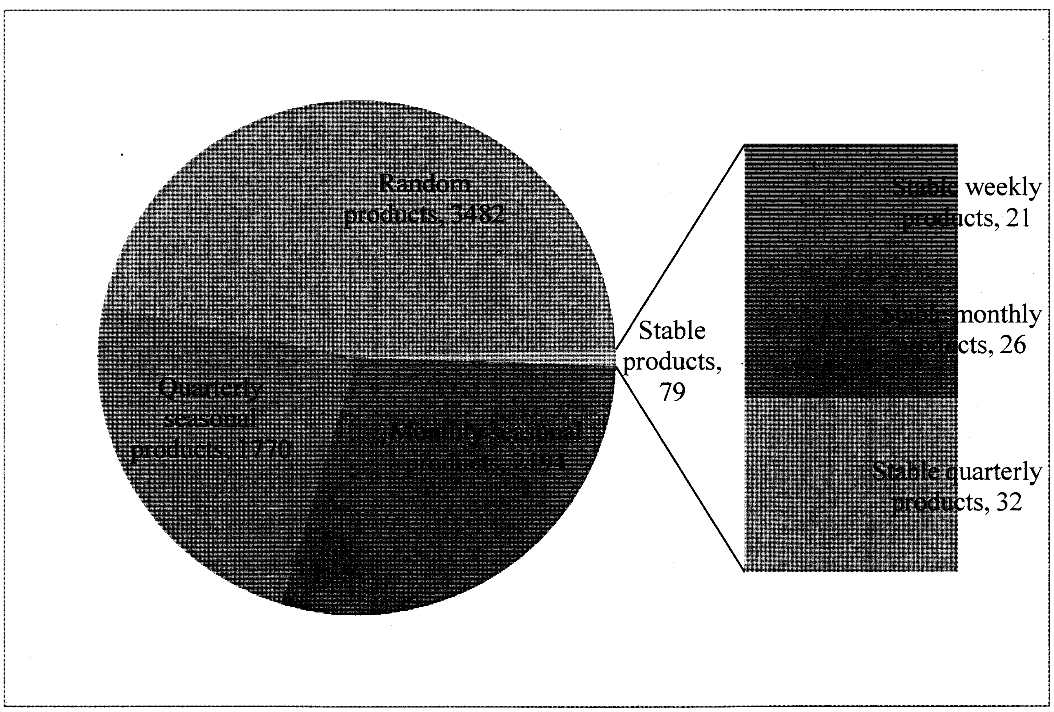


Figure 4-8: Product distribution in 2006

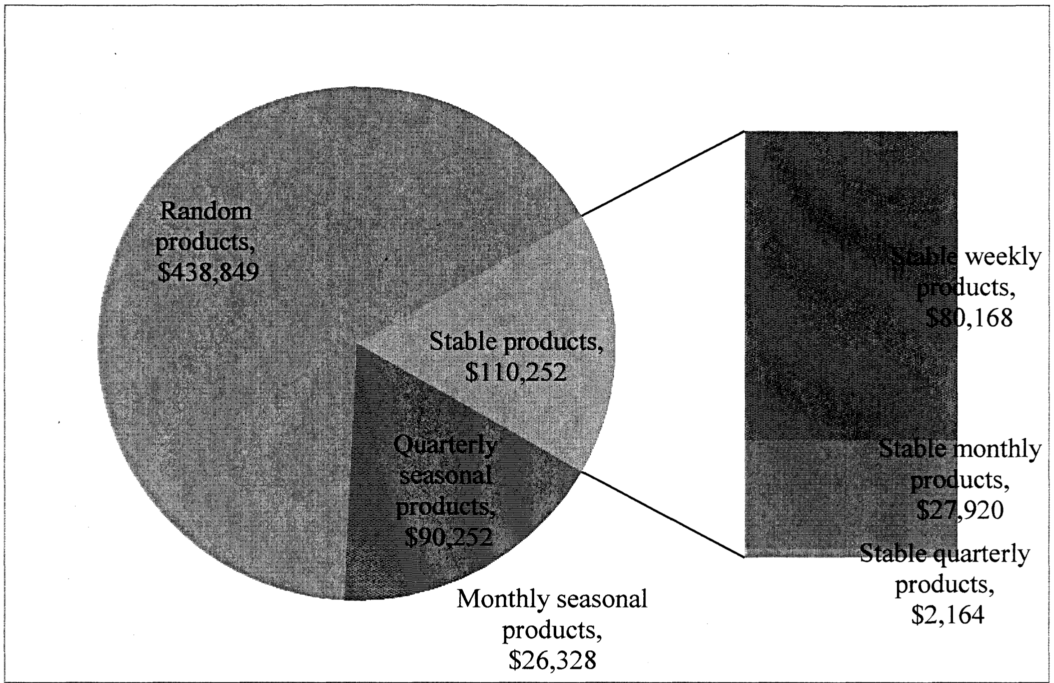


Figure 4-9: Profits distribution in 2005

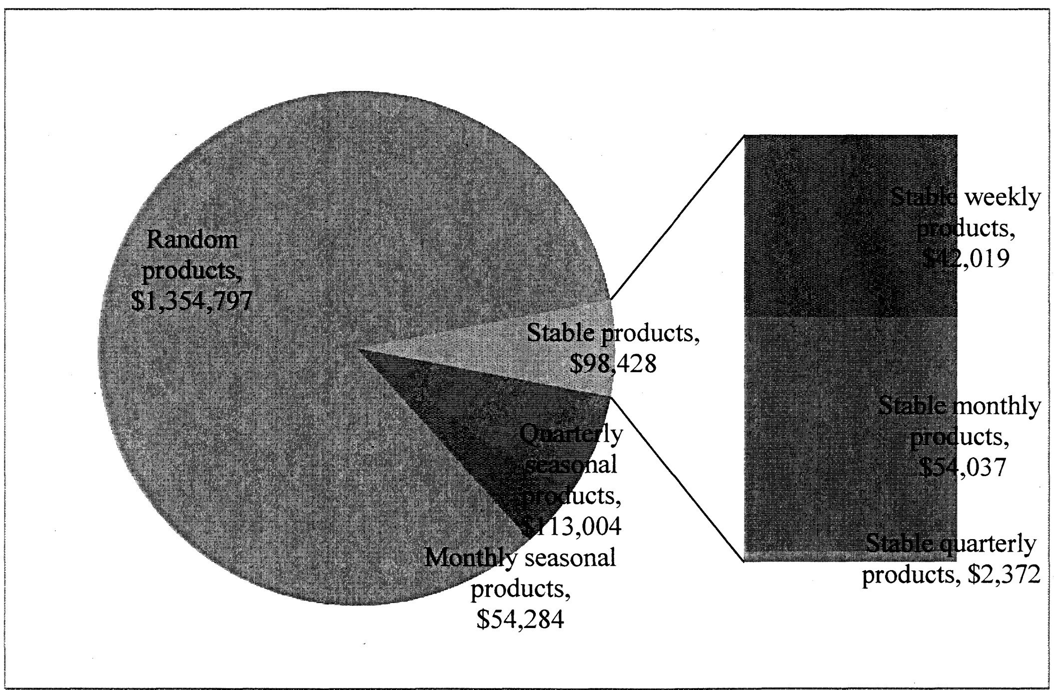


Figure 4-10: Profits distribution in 2006

Chapter 5

Inventory forecasting and business simulation

SPSS, the statistics software supported by IBM, is a well known data mining application. Based on product profiling and grouping results in Chapter 4, this study implements inventory forecasting experiments with SPSS. Time series prediction techniques, such as Simple Exponential Smoothing, Brown's Exponential Smoothing, Holt's Exponential Smoothing, Damp Trend Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA), are applied to forecast inventory demands. We aim to find the optimal solution for each product group and compare them with the generic optimal solution, which is the optimal prediction technique for the entire product set. In addition, a simulation program is proposed to generate business reports based on historical and predicted values. The business reports define optimal solutions from managerial points of review. We will also compare optimal solutions defined by statistical metrics and business reports.

5.1 Inventory forecasting

According to section 3.3.2.1, we create inventory forecasting models to predict inventory demands based on the proposed steps. Multiple time series prediction techniques are applied to create inventory forecasting models. Table 5-1 shows multiple forecasting results of time series prediction techniques. It lists monthly historical sales quantities for product P101 in 2005. Time series prediction techniques forecast quantities demanded for each month based on monthly sales quantities. For each month, predicted demands are listed according to prediction techniques.

Prediction period	Historical sales quantities	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average
January	0	1.31	1.21	1.24	1.35	1.5
February	2	1.11	1.28	0.72	1.22	1.5
March	2	1.24	2.15	1.14	1.83	1.5
April	2	1.36	2.71	1.44	2.28	1.5
May	2	1.45	2.89	1.66	2.51	1.5
June	2	1.53	2.76	1.82	2.53	1.5
July	3	1.60	2.43	1.92	2.41	1.5
August	2	1.81	2.68	2.39	2.72	1.5
September	0	1.84	2.35	2.32	2.50	1.5
October	2	1.57	0.82	1.47	1.21	1.5
November	1	1.63	0.70	1.66	1.03	1.5
December	0	1.54	0.40	1.40	0.66	1.5

Table 5-1: Multiple forecasting results based on product P101 in 2005

Mean absolute percentage error (MAPE) is one of the popularly used statistical evaluation metrics, discussed in section 2.3.1.3. It is applied to evaluate inventory forecasting models in this study. We define a time series prediction technique as the best-fit solution for a product if it has the lowest MAPE compared with other time series prediction techniques. Moreover, we compare the frequencies of the best-fit solutions' occurrence to define local optimal solution for each group. Similarly, generic optimal solutions are defined for the entire product set. Table 5-2 shows an example of MAPE comparisons for group 05M10. MAPE values of time series prediction techniques are listed accordingly based on each product. Based on MAPE comparison results, the best-fit solution, which has the lowest MAPE value, is identified on an item-by-item basis. For instance, MAPE values for product P000000000079 show that Damp Trend Exponential Smoothing is the best-fit solution since it has the lowest MAPE value (9.235403).

Product ID (2005)	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Best-fit solutions
P000000000079	9.249983	9.235475	9.249999	**9.235403**	9.243056	**Damp Trend Exponential Smoothing**
P01969250	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P020078110906	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P021245583127	7.592215	7.524391	7.616614	7.517398	7.548611	Damp Trend Exponential Smoothing
P02230106	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P02233116	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P023991000163	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P027434001472	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing
P030985006506	9.249983	9.235475	9.249999	9.235403	9.243056	Damp Trend Exponential Smoothing

Table 5-2: An example of MAPE comparisons for group 05M10

5.1.1 Generic (global) optimal solutions

A generic optimal solution is the time series prediction model that most frequently appeared to be the best-fit solution in the entire product set. It is also named a global optimal solution. We compared frequencies of the best-fit solutions' occurrences to determine generic optimal solutions. Two levels of inventory forecasting are performed to find generic optimal solutions at month and quarter level. Due to the large number of computations, inventory forecasting is not performed at week level. Table 5-3 illustrates generic optimal solutions at month and quarter level in 2005 and 2006. It compares frequencies of the best-fit solutions' occurrences. For example, in 2005, the occurrence frequency of Damp Trend Exponential Smoothing (1771) is the highest. That is, it is the best-fit solution for 1771 products. Thus, Damp Trend Exponential Smoothing is the generic optimal solution for the entire product set at month level in 2005. Autoregressive Integrated Moving Average (ARIMA) is the generic optimal solution at month level in 2006 since it is the best-fit solution for the most number of products (2360). The generic optimal solution at quarter level in 2005 is Holt's Exponential Smoothing, which is the same as the generic optimal solution in 2006.

Year	Level	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Generic optimal solutions
2005	Month	688	948	580	**1771**	1750	**Damp Trend Exponential Smoothing**
	Quarter	765	**1703**	1077	914	1370	**Holt's Exponential Smoothing**
2006	Month	958	1200	1085	1922	**2360**	**Autoregressive Integrated Moving Average**
	Quarter	991	**2009**	1586	1279	1788	**Holt's Exponential Smoothing**

Table 5-3: Generic solutions in 2005 and 2006

5.1.2 Local optimal solutions

Similarly, the best-fit prediction technique with the highest occurrence frequency in a group is the optimal solution for a given group. It is called the local optimal solution. Local optimal solutions may be different between groups. They may also be different from generic solutions.

Table 5-4 shows a sample distribution of local optimal solutions. The complete distribution of local optimal solutions is listed in Appendix Table A-1. Group names are defined based on categorizing processes in Chapter 4. For example, 05QStable is a stable quarterly group defined in 2005 quarterly stability analysis and 05M30 is a seasonal group defined in 2005 monthly seasonality analysis, where 3 and 0 are the group numbers in Tier-1 and Tier-2 analysis. Nature of the group, listed in the second column, indicates products' sales patterns in the group. Frequencies of the best-fit solutions' occurrences are compared in this table. The local optimal solution for each group is indicated based on the highest occurrence frequency criteria. Obviously, local optimal solutions are different between groups. Stable groups have no preferences in time series prediction techniques. That is, all inventory forecast models work equally well. In seasonal sales groups, only one time series prediction technique is defined as the local optimal solution. For example, the local optimal solution for group 05M30 is Brown's Exponential Smoothing since it has the highest occurrence frequency (130).

Group Name	Nature of the group: stability and seasonality	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Local Optimal Solutions
05QStable	Stable quarterly	**23**	**23**	**23**	**23**	**23**	**All the same**
05M30	January	5	6	**130**	32	6	**Brown's Exponential Smoothing**
05Q20	Quarter 1	0	0	**135**	0	0	**Brown's Exponential Smoothing**
06QStable	Stable quarterly	**32**	**32**	**32**	**32**	**32**	**All the same**

Table 5-4: The sample distribution of local optimal solutions

In addition, generic optimal solutions may not always be the same as local optimal solutions. The generic optimal solution at quarter level is Holt's Exponential Smoothing in 2005, as shown in Table 5-3. However, Table 5-4 indicates that Brown's Exponential Smoothing is the local optimal solution for the group 05Q20. There are 135 products in group 05Q20. Brown's Exponential Smoothing is the best-fit solution for all products within this group. We compared MAPEs of time series prediction models for group 05Q20 in Table 5-5. MAPEs associated with Brown's Exponential Smoothing (7.50035), which is the local optimal solution, is lower than Holt's Exponential Smoothing (7.541999), which is the generic optimal solution. Therefore, local optimal solutions forecast more accurately than generic optimal solutions. Table 5-6 shows comparison results of predicted quantity demands for P-4233149, which is a product in 05Q20. The predicted value of Brown's Exponential Smoothing (0.998598661) for Quarter 1 is the most reasonable value since its difference to the historical sales quantity (1) is the smallest. For Quarters 2, 3 and 4, because historical sales quantities are zeros, we set prediction errors as 10 for all prediction techniques to avoid invalid percentage error calculations. Moreover, statistical evaluation metrics, such as MAPE, may not always be good indicators. Managerial metrics, which evaluate inventory forecast based on managerial reviews, will be discussed in section 5.2.

Product ID	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Best-fit solutions
P-4233149	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P020078107555	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P020078108705	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P021718500293	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P021718500460	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P030985021257	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P033674136751	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P033674145371	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**
P033674606001	7.716615	7.541999	**7.50035**	7.541548	7.6875	**Brown's Exponential Smoothing**

Table 5-5: MAPE comparison results for group 05Q20

	Historical sales quantities	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average (ARIMA)
Quarter 1	1	0.133538532	0.832002838	**0.998598661**	0.833806838	0.25
Quarter 2	0	0.249244524	0.582381634	0.002800713	0.583426269	0.25
Quarter 3	0	0.215960776	0.107743714	-0.999996073	0.111020541	0.25
Quarter 4	0	0.187121691	-0.224564643	-0.001401826	-0.221603541	0.25

Table 5-6: Comparison results of predicted quantity demands for P-4233149 in 05Q20

5.1.3 Groups with strong sales patterns

Product profiling and clustering is essential to inventory forecasting. In Chapter 4, this study categorized products into reasonable groups based on their sales patterns. Each group is associated with a typical sales pattern. Groups, which are associated with very strong sales patterns, do not need prediction models. For example, products in the stable quarterly group were sold same number of times in each quarter. A simple solution is to keep these products at the stable quantity all quarters. Seasonal sales groups may also have very strong sales patterns. For instance, products in group 06M60 were only sold in July in 2006. Table 5-7 shows MAPE comparison results for group 06M60. MAPE values are high for all the prediction techniques. Table 5-8 shows comparison results of predicted values for product P0114-3. None of these prediction techniques work well for this case. A simpler inventory management, which orders a reasonable amount in its sales month (July), will serve very well. For the rest of the year, do not carry this product.

Product ID	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	**Autoregressive Integrated Moving Average (ARIMA)**	**Best-fit solutions**
P0114-3	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P02185253	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P02233206	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P02233899	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P030985004687	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P030985007800	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P033674000724	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P036923000612	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P036923029149	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**
P036923291201	9.248119	9.247283	9.243259	9.249918	**9.243056**	**ARIMA**

Table 5-7: MAPE comparison results for group 06M60

Product P0114-3 in 06M60	Historical sales quantities	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average
January	0	0.027829	0.030392311	0.098211103	0.00099303	0.083333333
February	0	0.026874	0.030864257	0.095094556	0.00099204	0.083333333
March	0	0.025952	0.031289138	0.092073992	0.00099104	0.083333333
April	0	0.025061	0.03167163	0.089146518	0.00099005	0.083333333
May	0	0.024201	0.032015944	0.086309328	0.00098906	0.083333333
June	0	0.023371	0.032325873	0.083559697	0.00098807	0.083333333
July	1	0.022569	0.032604833	0.080894983	0.00098708	0.083333333
August	0	0.056109	0.132250384	0.13088269	0.001987	0.083333333
September	0	0.054184	0.122602191	0.126307484	0.00198501	0.083333333
October	0	0.052325	0.113912297	0.121882369	0.00198303	0.083333333
November	0	0.050529	0.106085503	0.117602617	0.00198104	0.083333333
December	0	0.048795	0.099036065	0.113463643	0.00197906	0.083333333

Table 5-8: Comparison results of predicted values for product P0114-3

This study applied time series clustering models to discover products' sales patterns and categorize them into reasonable groups, such as stable groups, single-month selling groups, two-months selling groups, single-quarter selling groups and two-quarters selling groups. Defining an appropriate prediction level (period) based on sales patterns is critical for inventory forecasting. For example, products in single-quarter selling group should be properly predicted at Quarter level. Products in 05Q03 have a two-quarters selling pattern in Quarters 1 and 2. Tables 5-9 and 5-10 illustrate MAPE comparison results for 05Q03 at month and quarter level. Obviously, MAPE results at month level are higher than MAPE results at quarter level. The local optimal solution is defined differently at these two levels. The lower the MAPE is, the better the prediction model is. Thus, Autoregressive Integrated Moving Average, the local optimal solution defined at quarter level, outperforms Damp Trend Exponential Smoothing, the local optimal solution defined at month level.

Product ID	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Best-fit solutions
P018788801603	6.902311	6.891110	6.906179	6.908367	6.878472	**ARIMA**
P02233531	7.690860	7.652563	7.698958	7.652772	7.668981	**Holt's Exponential Smoothing**
P030985004953	7.673619	7.629887	7.693950	7.624153	7.687500	**Damp Trend Exponential Smoothing**
P051381311353	7.673619	7.629887	7.693950	7.624153	7.687500	**Damp Trend Exponential Smoothing**
P058854490218	6.855045	6.821750	6.841989	6.821234	6.822222	**Damp Trend Exponential Smoothing**

Table 5-9: MAPE comparison results based on the month level prediction for group 05Q03

Product ID	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average (ARIMA)	Best-fit solutions
P018788801603	2.798991	2.875208	3.055680	2.874711	2.770833	**ARIMA**
P02233531	2.798991	2.875208	3.055680	2.874711	2.770833	**ARIMA**
P030985004953	5.284730	5.292634	5.311855	5.293069	5.218750	**ARIMA**
P051381311353	5.284730	5.292634	5.311855	5.293069	5.218750	**ARIMA**
P058854490218	3.003959	3.350442	3.416403	3.349475	3.166667	**Simple Exponential Smoothing**

Table 5-10: MAPE comparison results based on the quarter level prediction for group 05Q03

5.1.4 Cross-year comparison

Product profiling and clustering and inventory forecast are performed on a yearly basis. The majority of products may actually be sold in many years. In the retail chain data set, there are 5102 products sold in 2005 and 2006. The retail chain is a specialty store. It changes inventories based on current trends. Many substitute products are imported into stores from year to year. For example, Vitamin C is a common health care product. Many brands of Vitamin C are available in the market. The store switched vitamin C from brand to brand. This is an effective marketing strategy to keep attracting customers, who look for the latest trend. The retail chain owned one store in 2005 and expanded to three stores in 2006. Compared to the revenue in 2005, the revenue in 2006 is doubled. A cross-year comparison illustrates product sales patterns more clearly. Products, who share same sales patterns in 2005 and 2006, may have the same sales behaviours in 2007. This could help inventory decision makers to pre-define products into reasonable groups. However, products may not always share same sales patterns in different years. This could be caused by many reasons: i) products have short business lives, ii) products were sold firstly in the middle of a year, iii) customers' loyalty to the product has changed, it could be increased or decreased, iv) competition from alternative products and v) macroeconomic factors, such as employment rate and bank rate, that affect customer consumption power. Statistically, small sales quantity changes on products, which were sold at a low quantity for a year, could also dramatically change sales patterns. For example, a product was only sold one time in December 2005, so it is a December selling product. In 2006, if it was only sold one time in January, it will be defined as a January selling product. Its sales pattern changed significantly from 2005 to 2006. Table 5-11 shows cross year product comparison results for the stable monthly and weekly group 05MW in 2005. There are 119 products with the same sales pattern in 05MW. In 2006, 11 of them still have the same sales pattern, which is stably sold at month and week level. Table 5-12 shows stable monthly and weekly products that were sold in 2005 and 2006. Since products, listed in Table 5-12, shared the same sales patterns in 2005 and 2006, it would be appropriate management strategy if these products are categorized in to stable monthly and weekly sales group in 2007. In addition, there are seven products with different sales patterns. Thus, they are categorized into groups 06M20, 06Q11, 06Q20, 06Q21, 06Q22, 06Q23 and 06Q63, respectively. The rest of 101 products are random selling products in group 06Random. This means that many stable monthly and weekly products in 2005 have changed their sales behaviours to random selling in 2006.

Group name	Nature of the group: stability and seasonality	Number of products
05MW	Stable monthly and weekly	119
06M20	December	1
06MW	**Stable monthly and weekly**	**11**
06Q11	Quarters 4 and 3	1
06Q20	Quarter 4	1
06Q21	Quarters 4 and 1	1
06Q22	Quarters 4 and 3	1
06Q23	Quarters 4 and 1	1
06Q63	Quarters 1 and 2	1
06Random	Random	101

Table 5-11: Cross-year comparison results for group 05MW

Product ID	Year	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
P068958011219	2005	6	18	23	9	8	3	6	9	7	19	10	11
	2006	60	21	14	24	30	22	21	15	16	17	17	77
P624917060027	2005	6	6	9	9	8	11	14	15	19	14	9	13
	2006	33	8	15	23	77	24	17	24	28	18	16	11
P693749015017	2005	22	31	23	34	28	32	30	28	24	34	26	30
	2006	91	48	49	58	54	57	50	51	53	60	52	65
P777672011954	2005	18	87	134	48	48	31	38	16	36	112	47	71
	2006	356	58	128	77	73	110	382	83	82	94	61	326
P790011040033	2005	7	7	4	10	8	12	4	10	7	6	9	9
	2006	9	6	14	13	10	7	29	11	11	14	12	13
P790011060123	2005	26	27	29	26	30	14	14	20	4	9	22	37
	2006	133	36	46	62	62	83	70	74	56	65	70	81
P838766005829	2005	15	6	39	11	10	28	38	37	20	24	16	51
	2006	32	53	44	53	50	51	43	38	27	96	53	60
P4004148047527	2005	3	8	7	4	4	5	2	5	0	2	2	4
	2006	21	13	18	14	9	16	16	10	32	15	17	17
P631257355553	2005	3	1	3	1	2	2	6	4	0	2	1	2
	2006	8	8	16	10	7	8	9	7	4	5	8	4
P631257535313	2005	11	7	62	0	2	2	22	32	12	4	4	27
	2006	27	11	21	13	44	30	19	12	34	8	31	139

Table 5-12: Stable monthly and weekly products in 2005 and 2006

5.2 Business simulation

Inventory forecasting is critical to inventory management. Statistical measurements of forecast variability are important to inventory decision making. In addition, managerial metrics provide decision makers with inventory management reports. According to Gardner's (1990) total variable cost formula in inventory control system, discussed in section 2.3.2, this study proposes a simulation program to simulate business operations based on historical sales records and predicted sales quantities. This program evaluates inventory forecast results from a managerial point of view. It calculates the length of shortage periods, counts the product quantity left at the end of the year and generates cost management reports. The length of shortage periods is the summation of periods when customer loyalty is lost due to the product's unavailability. The quantity left at the end of the year is the number of leftovers based on business simulation with predicted values. Since replenishment costs do not change significantly with inventory forecast results, only carrying costs and shortage costs are included in cost management reports. Hence, the total variable cost, denoted by T, is defined as Equation 5-1:

$$T = ICQ + (P\text{-}C)\ RL \qquad \text{Equation 5-1}$$

Where,

I = interest rate, defined as 5% in this study

C = unit purchase cost of the item

Q = quantity of the item in stock.

P = unit selling price of the item.

R = number of units required by customers.

L = customer loyalty loss factor, defined as 1.1 in this study.

ICQ = carrying cost due to over-ordered inventories in stock.

(P-C)RL = shortage cost due to the unavailability of inventories.

Table 5-13 shows MAPE comparison results based on product P114. It also includes historical and predicted values. Since the MAPE value of Damp Trend Exponential Smoothing (0.278803976) is the smallest, statistically, the best-fit solution for P114 is Damp Trend Exponential Smoothing. Table 5-14 illustrates a management report of business simulation based on product P114. Similar to Table 5-13, historical sales quantities and predicted values are listed according to months. In addition, the length of shortage periods, quantities left at the end of the year and total variable cost, associated with time series techniques, are shown accordingly. The business simulation program evaluates time series prediction models from a managerial angle. The length of shortage periods and quantities left are good managerial indicators for inventory management. Total variable cost is one of the most important managerial indicators for inventory forecast evaluations. Cost minimization is the key to improve business performance. Comparisons of three managerial indicators among time series prediction techniques shows that Autoregressive Integrated Moving Average has the shortest shortage period (1 month), the optimal quantities left at the end of the year (0.00) and the lowest total variable cost ($5.44). Thus, inventory management with predicted sales quantities based on Autoregressive Integrated Moving Average leads to the best business performance. Although the statistical metric (MAPE) indicated that Damp Trend Exponential Smoothing is the best-fit solution for P114, our managerial report proved that Autoregressive Integrated Moving Average is better in terms of business operation. Cost management is more critical than statistical metrics.

	Historical sales quantity	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average
January	8	7.30	5.64	10.19	5.18	10.83
February	3	7.60	6.53	9.74	6.09	10.83
March	10	5.64	7.32	6.58	6.98	10.83
April	9	7.50	8.20	8.56	7.88	10.83
May	6	8.14	9.08	9.15	8.79	10.83
June	10	7.23	9.87	7.77	9.68	10.83
July	15	8.41	10.70	9.14	10.58	10.83
August	9	11.22	11.63	12.71	11.49	10.83
September	17	10.27	12.45	11.40	12.38	10.83
October	14	13.15	13.41	14.99	13.30	10.83
November	13	13.51	14.32	15.34	14.21	10.83
December	16	13.29	15.20	14.87	15.12	10.83
MAPE		0.350772068	0.283000046	0.425840559	**0.278803976**	0.475100341

Table 5-13: MAPE comparison results based on product P114

	Historical sales quantities	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average
January	8	7.30	5.64	10.19	5.18	10.83
February	3	7.60	6.53	9.74	6.09	10.83
March	10	5.64	7.32	6.58	6.98	10.83
April	9	7.50	8.20	8.56	7.88	10.83
May	6	8.14	9.08	9.15	8.79	10.83
June	10	7.23	9.87	7.77	9.68	10.83
July	15	8.41	10.70	9.14	10.58	10.83
August	9	11.22	11.63	12.71	11.49	10.83
September	17	10.27	12.45	11.40	12.38	10.83
October	14	13.15	13.41	14.99	13.30	10.83
November	13	13.51	14.32	15.34	14.21	10.83
December	16	13.29	15.20	14.87	15.12	10.83
Shortage period		10 months	9 months	2 months	11 months	**1 month**
Quantities left		(16.74)	(5.63)	0.43	(8.33)	**(0.00)**
Total variable cost		$759.76	$308.15	$25.65	$498.55	**$5.44**

Table 5-14: A sample report of business simulation based on product P114

Similarly, local optimal solutions may be defined differently by statistical metrics and by managerial metrics. Total variable cost is the key to evaluate inventory forecast results. To identify local optimal solutions, this study compared total variable costs in product groups. Table 5-15 illustrates sample distributions of local optimal solutions based on business simulation reports. The complete distribution of total variable costs based on product groups is listed in Appendix Table A-2. The chosen groups are the same as the groups in Table 5-4. Therefore, we could compare local optimal solutions indicated by MAPE and business simulation reports. The bolded and highlighted values are the lowest total variable costs for these groups. They also indicate that corresponding time series prediction techniques, such as Damp Trend Exponential Smoothing and Holt's Exponential Smoothing are the local optimal solutions, respectively. Results are different than local optimal solutions defined with MAPE in Table 5-4. MAPE and business simulation reports point at the same local optimal solutions for groups 05QStable, 06QStable and 06Q32. Again, business simulation reports provide more practical results for inventory management. Therefore, inventory decisions should be made according to business simulation reports.

Group Name	Nature of the group: stability and seasonality	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average	Local Optimal Solutions
05QStable	Stable quarterly	$0	$0	$0	$0	$0	All the same
05M30	January	$5,336.21	$2,710.38	$6,183.04	**$2,404.06**	$6,261.19	**Damp Trend Exponential Smoothing**
05Q20	Quarter 1	$5,314.01	$477.11	$4,863.42	**$472.28**	$3,776.60	**Damp Trend Exponential Smoothing**
06QStable	Stable quarterly	$0	$0	$0	$0	$0	All the same

Table 5-15: Sample distributions of local optimal solutions based on business simulation reports

5.3 Summary of inventory forecasting and business simulation

5.3.1 Inventory forecasting and time series prediction

Many commonly used time series prediction techniques were applied to forecast inventory demands in this study. We compared MAPE evaluation results between time series prediction models and identified the best-fit solution for each product, the local optimal solution for each group and the generic optimal solution for the entire product set in each year at month and quarter level.

In 2005, generic optimal solutions at month and quarter level are Damp Trend Exponential Smoothing and Holt's Exponential Smoothing. Local optimal solutions for product groups in 2005 are identified as below:

- Simple Exponential Smoothing for 8 groups
- Holt's Exponential Smoothing for 10groups
- Brown's Exponential Smoothing for 8 groups
- Damp Trend Exponential Smoothing for 9 groups
- Autoregressive Integrated Moving Average for 17 groups
- Time series prediction techniques worked equally well for 1 group, which is the stable monthly and weekly group

In 2006, generic optimal solutions at month and quarter level are Autoregressive Integrated Moving Average and Holt's Exponential Smoothing. Local optimal solutions for product groups in 2006 are identified as below:

- Simple Exponential Smoothing for 2 groups

- Holt's Exponential Smoothing for 14 groups
- Brown's Exponential Smoothing for 8 groups
- Damp Trend Exponential Smoothing for 9 groups
- Autoregressive Integrated Moving Average for 18 groups
- Time series prediction techniques worked equally well for 1 group, which is the stable monthly and weekly group

Moreover, we compared generic optimal solutions to local optimal solutions in specific groups. Local optimal solutions outperformed generic optimal solutions based on the lower MAPE criterion. Furthermore, statistical metric like MAPE may not always be good indicators. Product groups with strong sales patterns may not need inventory forecasting. It would be a lot simpler to control seasonal inventories using proposed inventory management strategies.

5.3.2 Business simulation

From managerial points of view, the goal of modern inventory management is to minimize costs of carrying inventory without ever running out of products. A simple simulation program was proposed to evaluate inventory forecasting results from the business management angle. It calculated the length of shortage periods, the quantity of inventory remained in stock and the total variable cost and generated business reports. Hence, the total variable cost is the sum of carrying costs and shortage costs. The carrying cost is calculated using:

- interest rate, defined as 5% in this study
- unit purchase cost of the item

- quantity of the item in stock

The shortage cost is calculated using:

- unit selling price of the item
- number of units required by customers
- customer loyalty loss factor, defined as 1.1 in this study

The total variable cost is the key business indicator to identify the best-fit solution for each product and the local optimal solutions for product groups. We compared total variable costs to identify the local optimal solutions for each group. Local optimal solutions for product groups in 2005 are identified as below:

- Simple Exponential Smoothing for 4 groups
- Holt's Exponential Smoothing for 9 groups
- Brown's Exponential Smoothing for 6 groups
- Damp Trend Exponential Smoothing for 11 groups
- Autoregressive Integrated Moving Average for 23 groups

Local optimal solutions for product groups in 2006 are identified as below:

- Simple Exponential Smoothing for 5 groups
- Holt's Exponential Smoothing for 10 groups
- Brown's Exponential Smoothing for 5 groups
- Damp Trend Exponential Smoothing for 8 groups
- Autoregressive Integrated Moving Average for 24 groups

For each group, the local optimal solution may be defined differently by statistical metrics and managerial metrics. Based on comparison of results, Cost/benefit analysis based on the proposed simulation may be more relevant to the inventory manager than statistical measure such as MAPE.

Chapter 6

Conclusions

6.1 Summary and Conclusions

Inventory management (IM), as an essential business issue, plays a significant role in improving business performance. In inventory management, a number of objectives are of interest to inventory managers. These include maximizing profits (with or without discounts), rates of return on investment, the chance of survival, minimizing cost (with or without discounts), ensuring flexibility of operations and determining feasible solutions. Efficient and effective inventory management increases inventory accuracy, automates order process and optimizes business productivity.

Inventory forecasting leads a critical path to support decision makings in inventory management. Many researchers have paid significant attention to this area. However, most, if not all, inventory forecasting studies performed time series prediction without time series clustering techniques, which conveniently discover data distribution and patterns. In this study, we applied time series clustering techniques, such as Expectation Maximization (EM) and K-Means algorithms, to categorize products into appropriate groups.

The stability analysis was performed based on products' sales patterns at three levels: week, month and quarter. Two tiers of clustering analysis were applied for each level. Stable groups at week, month and quarter level and an unstable group were identified. Moreover, the seasonality analysis was performed on the unstable group at two levels:

month and quarter. Again, two tiers of clustering analysis were applied for each level. Seasonal groups were identified based on time series clustering results. According to product analyses results, 5735 products were categorized into 54 groups including 3 stable groups, 50 seasonal groups and 1 random group in 2005. In 2006, 7525 products were categorized into 53 groups including 3 stable groups, 49 seasonal groups and 1 random group.

Inventory forecasting predicts the future inventory demands based on products' historical sales quantities. Many time series prediction techniques, such as Regression, Exponential Smoothing, Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks, are commonly used in inventory forecasting. This study applied Simple Exponential Smoothing, Holt's Exponential Smoothing, Brown's Exponential Smoothing, Damped Trend Exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) to forecast inventory demands based on historical sales quantities.

Mean Absolute Percentage Error (MAPE), as a common statistical evaluation measure, was used to evaluate inventory forecast results. Generic optimal forecasting solutions and local optimal forecasting solutions were identified based on MAPE comparison results. Local optimal solutions outperformed generic optimal solutions in product groups. Furthermore, products with strong sales patterns may not need forecasting solution. For examples, products with high stability patterns can be kept at the same level in all periods. Products with strong seasonality patterns can be managed with the proposed inventory management strategies:

(i) Carry very few seasonal products in their off-sales periods. The inventory in off-season should be based on quantities from previous year sales during off-season.

(ii) Order a lot of seasonal products in their sales periods. The size of order should be based on quantities from previous year sales during the same season.

In addition, statistical evaluation metrics may not always be good indicators of inventory forecasting. A simple simulation program was proposed to simulate business operations with predicted sales quantities, historical sales quantities, prices and costs. This study developed a total variable cost formula to perform cost/benefit analysis. It summed up carrying costs, which are due to over stocked inventories, and shortage costs, which are due to unavailability of inventories. Business management reports including the length of shortage periods, remaining quantities and total variable costs were generated by the simulation program. Local optimal solutions for product groups were also identified based on business reports. For each group, the local optimal solution may be defined differently by statistical metrics (MAPE) and managerial metrics (cost/benefit analysis). Based on comparison results, managerial metrics - cost/benefit analysis may be more relevant to inventory decision makings than statistical metrics - MAPE.

6.2 Future directions

This study applied data mining techniques, such as time series clustering and time series prediction, in inventory management. Business simulation program was created to study cost and benefits of inventory forecasting. Potential future directions of this study are discussed below.

1. Study the use of product profiling and time series clustering analysis with soft clustering techniques, such rough and fuzzy clustering. Time series clustering

algorithms, such as Expectation Maximization (EM) and K-Means were applied to profile products based on their sales patterns. However, the results suggested that the clusters may overlap. Soft clustering techniques, such as fuzzy clustering and rough clustering, that allow for overlapping clusters may provide better clustering schemes.

2. Consider hybrid prediction techniques for inventory forecasting. Since product groups can be categorized with combinations of multiple sales patterns based on soft clustering techniques, hybrid prediction techniques should be applied to inventory forecasting. That is, inventory forecasting should be performed based on combinations of multiple optimal solutions. The weight of optimal solutions should be defined based on product's fuzzy membership in a given cluster. Inventory forecasting with other time series prediction techniques, such as artificial neural networks, may also be a potential research direction.
3. Study the effect of discounts on demand forecasting. The retailer offers discounts to attract customers or to reduce the inventory. Sales quantities increase significantly as a result of the discount. A study of demand based on different types of discounts should be part of the inventory management strategy. The goal of such a study will be to reduce the amount of products that have to be placed on clearance, and to maximize the impact of discounts intended to attract customers.
4. Refine the simulation program to consider more sophisticated business parameters. The simulation program proposed in this study used a limited number of parameters in calculating cost and benefits. Business operations were simulated based on historical sales records and predicted sales quantities. In the future research, we may refine the simulation program so that it can process more

business information, such as inventory orders and special sales events. That is, the refined simulation program should be able to simulate sophisticated business operations, which include cost of ordering products, delivering products, cost of maintaining products on shelves, and selling products with or without discounts.

Appendix

Group name	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average (ARIMA)	Local optimal solutions
05M00	0	105	0	1	0	Holt's Exponential Smoothing
05M01	2	6	1	8	6	Damp Trend Exponential Smoothing
05M02	0	0	0	0	4	ARIMA
05M03	0	0	2	0	0	Brown's Exponential Smoothing
05M04	0	0	0	0	2	ARIMA
05M10	0	1	2	295	0	Damp Trend Exponential Smoothing
05M11	0	0	0	0	5	ARIMA
05M12	0	0	2	0	2	Brown's Exponential Smoothing
05M14	1	0	1	2	3	ARIMA

05M20	2	1	0	17	0	Damp Trend Exponential Smoothing
05M21	0	0	0	154	0	Damp Trend Exponential Smoothing
05M22	28	3	0	4	1	Simple Exponential Smoothing
05M24	0	0	0	3	0	Damp Trend Exponential Smoothing
05M30	5	6	130	32	6	Brown's Exponential Smoothing
05M31	0	8	0	2	0	Holt's Exponential Smoothing
05M32	0	0	0	2	5	ARIMA
05M33	0	0	13	1	3	Brown's Exponential Smoothing
05M34	0	0	0	3	0	Damp Trend Exponential Smoothing
05M41	0	4	0	0	5	ARIMA
05M42	0	0	0	0	6	ARIMA
05M43	0	0	1	0	89	ARIMA
05M44	0	1	0	0	10	ARIMA
05M50	3	7	7	124	115	Damp Trend Exponential Smoothing
05M52	0	1	2	4	96	ARIMA

05M91	1	1	0	1	75	ARIMA
05M92	0	9	3	11	2	Damp Trend Exponential Smoothing
05M94	1	2	12	1	11	Brown's Exponential Smoothing
05MW	32	21	14	23	29	Simple Exponential Smoothing
05Q00	3	2	1	212	0	Damp Trend Exponential Smoothing
05Q01	2	4	1	0	38	ARIMA
05Q02	5	0	0	0	8	ARIMA
05Q03	5	1	0	0	49	ARIMA
05Q04	6	2	4	4	0	Simple Exponential Smoothing
05Q10	0	41	0	5	0	Holt's Exponential Smoothing
05Q11	5	24	0	6	8	Holt's Exponential Smoothing
05Q12	5	29	9	1	7	Holt's Exponential Smoothing
05Q13	0	120	2	0	0	Holt's Exponential Smoothing
05Q14	2	13	1	0	4	Holt's Exponential Smoothing
05Q20	0	0	135	0	0	Brown's Exponential Smoothing

05Q21	15	1	1	1	0	Simple Exponential Smoothing
05Q22	16	0	1	2	2	Simple Exponential Smoothing
05Q23	10	1	1	0	0	Simple Exponential Smoothing
05Q24	3	7	52	17	7	Brown's Exponential Smoothing
05Q30	23	137	8	45	0	Holt's Exponential Smoothing
05Q32	9	2	80	9	12	Brown's Exponential Smoothing
05Q41	49	2	0	9	12	Simple Exponential Smoothing
05Q43	6	1	18	3	101	ARIMA
05Q44	3	1	4	3	43	ARIMA
05Q53	4	47	0	17	39	Holt's Exponential Smoothing
05Q54	73	32	7	35	2	Simple Exponential Smoothing
05Q61	4	71	0	0	5	Holt's Exponential Smoothing
05QStable	23	23	23	23	23	All the same
05Random	408	486	347	405	647	ARIMA
06M00	0	0	0	76	0	Damp Trend Exponential Smoothing

06M02	2	2	1	0	11	ARIMA
06M03	0	0	0	0	12	ARIMA
06M04	0	0	11	0	3	Brown's Exponential Smoothing
06M11	0	1	12	1	6	Brown's Exponential Smoothing
06M12	0	1	1	0	107	ARIMA
06M14	0	20	1	4	2	Holt's Exponential Smoothing
06M20	16	18	25	210	59	Damp Trend Exponential Smoothing
06M21	1	1	0	90	0	Damp Trend Exponential Smoothing
06M22	3	20	1	13	6	Holt's Exponential Smoothing
06M23	3	0	23	1	4	Brown's Exponential Smoothing
06M24	0	10	0	7	3	Holt's Exponential Smoothing
06M30	0	21	0	0	0	Holt's Exponential Smoothing
06M31	1	3	1	0	4	ARIMA
06M33	0	14	0	4	7	Holt's Exponential Smoothing
06M34	1	1	1	135	0	Damp Trend Exponential Smoothing

06M50	3	6	492	17	9	Brown's Exponential Smoothing
06M51	2	3	1	1	3	Holt's Exponential Smoothing
06M52	2	2	1	7	1	Damp Trend Exponential Smoothing
06M60	0	0	0	0	84	ARIMA
06M61	0	2	0	0	0	Holt's Exponential Smoothing
06M62	0	0	0	0	3	ARIMA
06M63	0	0	1	1	4	ARIMA
06M70	1	3	1	119	5	Damp Trend Exponential Smoothing
06M72	0	14	11	20	107	ARIMA
06M84	4	10	11	10	102	ARIMA
06M91	1	0	0	102	0	Damp Trend Exponential Smoothing
06M92	1	6	0	14	3	Damp Trend Exponential Smoothing
06M94	2	1	1	4	15	ARIMA
06MW	9	9	2	12	15	ARIMA
06Q01	2	3	6	0	66	ARIMA

06Q02	15	10	3	23	47	ARIMA
06Q03	5	0	3	4	41	ARIMA
06Q04	41	8	0	20	7	Simple Exponential Smoothing
06Q11	28	102	10	28	1	Holt's Exponential Smoothing
06Q20	0	188	12	0	0	Holt's Exponential Smoothing
06Q21	1	6	6	1	10	ARIMA
06Q22	14	20	16	5	15	Holt's Exponential Smoothing
06Q23	3	26	7	0	6	Holt's Exponential Smoothing
06Q24	0	29	1	14	7	Holt's Exponential Smoothing
06Q32	4	81	6	31	5	Holt's Exponential Smoothing
06Q33	11	10	58	3	29	Brown's Exponential Smoothing
06Q40	22	3	0	8	67	ARIMA
06Q41	24	61	3	0	10	Holt's Exponential Smoothing
06Q42	11	0	2	46	5	Damp Trend Exponential Smoothing
06Q44	2	0	1	0	41	ARIMA

06Q60	10	1	26	4	2	Brown's Exponential Smoothing
06Q62	98	10	0	12	0	Simple Exponential Smoothing
06Q63	20	22	71	10	37	Brown's Exponential Smoothing
06Q64	0	0	98	0	1	Brown's Exponential Smoothing
06QStable	32	32	32	32	32	All the same
06Random	552	841	434	695	904	ARIMA

Table A-1: The complete distribution of local optimal solutions based on MAPE

Group name	Simple Exponential Smoothing	Holt's Exponential Smoothing	Brown's Exponential Smoothing	Damp Trend Exponential Smoothing	Autoregressive Integrated Moving Average (ARIMA)	Local optimal solutions
05M00	1223.5453	1547.1068	236.67043	1595.4607	262.9794	Brown's Exponential Smoothing
05M01	1004.54364	1631.5512	235.60826	1487.4812	105.3909	ARIMA
05M02	103.879921	162.9856	13.144572	163.04367	17.28542	Brown's Exponential Smoothing
05M03	15.0874733	56.361736	3.9392616	49.196107	2.145833	ARIMA
05M04	26.4677263	40.104952	2.385105	41.053689	3.94349	Brown's Exponential Smoothing
05M10	4136.48904	181.22617	4262.13	199.84477	94.96432	ARIMA
05M11	49.9588396	35.654473	119.54558	34.89925	19.65339	ARIMA
05M12	539.003414	420.56146	420.44043	420.56333	9.218605	ARIMA
05M14	639.150427	273.75598	501.08881	269.88232	69.93119	ARIMA
05M20	881.648131	272.36961	1278.4306	244.74717	1321.128	Damp Trend Exponential Smoothing
05M21	2915.08339	963.2043	3017.9539	841.47741	2359.684	Damp Trend Exponential Smoothing

05M22	578.39473	586.69217	3028.4617	1329.2367	4320.754	Simple Exponential Smoothing
05M24	219.93918	31.803053	254.34941	31.701619	154.4559	Damp Trend Exponential Smoothing
05M30	5336.2121	2710.3889	6183.0455	2404.0647	6261.197	Damp Trend Exponential Smoothing
05M31	304.418372	59.622635	369.87729	58.972868	263.6944	Damp Trend Exponential Smoothing
05M32	364.414646	114.53921	427.21717	163.22328	223.6377	Holt's Exponential Smoothing
05M33	44.9082556	5414.9425	5789.6623	1149.7016	6050.146	Simple Exponential Smoothing
05M34	100.229831	32.377885	190.156	32.232192	165.4813	Damp Trend Exponential Smoothing
05M41	542.131807	425.55452	764.15655	377.27124	250.4934	ARIMA
05M42	235.733173	257.00057	321.68256	190.48581	103.4501	ARIMA
05M43	1452.86883	1432.7979	344.12593	1863.7987	493.3272	Brown's Exponential Smoothing
05M44	544.775533	444.65231	739.67461	485.06128	264.485	ARIMA
05M50	5453.32527	7686.0972	1513.6864	7950.4077	1119.12	ARIMA
05M52	1903.62616	2230.9325	321.98993	2309.2994	463.9859	Brown's Exponential Smoothing
05M91	1322.43325	1262.7962	1713.7652	1161.2699	646.4542	ARIMA
05M92	1480.24019	527.84581	1673.7223	628.88526	1004.874	Holt's Exponential Smoothing

05M94	1563.35788	1764.8173	401.78372	1813.5672	252.261	ARIMA
05MW	32678.49	20721.326	29726.534	21759.084	25252.07	Holt's Exponential Smoothing
05Q00	1671.99225	950.97659	2463.0321	954.49154	1113.119	Holt's Exponential Smoothing
05Q01	1970.87401	1814.4202	2689.6134	1785.8315	708.4072	ARIMA
05Q02	1415.59169	170.01624	1494.7486	479.96405	442.6893	Holt's Exponential Smoothing
05Q03	2546.9521	253.66614	2685.0166	251.10704	1536.177	Damp Trend Exponential Smoothing
05Q04	2015.17467	729.81209	2519.2848	695.26901	637.9329	ARIMA
05Q10	3151.50593	1247.1438	2140.2535	1286.8995	79.01281	ARIMA
05Q11	1098.00763	3249.8681	711.21143	3209.0816	103.3805	ARIMA
05Q12	3809.98844	2834.7201	2909.0652	2733.5161	120.2764	ARIMA
05Q13	1776.63977	2048.509	3018.7673	2023.1756	157.7656	ARIMA
05Q14	340.428231	519.51485	3510.3692	519.40386	221.0211	ARIMA
05Q20	5314.01686	477.11308	4863.4293	472.28461	3776.609	Damp Trend Exponential Smoothing
05Q21	377.132375	129.35334	1048.9176	291.88631	522.884	Holt's Exponential Smoothing
05Q22	2671.5385	193.63147	2509.8736	192.30358	1722.557	Damp Trend Exponential Smoothing

05Q23	839.232846	117.21793	1180.4035	129.35326	596.1195	Holt's Exponential Smoothing
05Q24	1818.37507	374.24776	5347.7666	363.86434	6071.005	Damp Trend Exponential Smoothing
05Q30	11414.9705	6318.4555	2807.3375	6342.3411	314.7759	ARIMA
05Q32	948.284831	869.09093	1858.0819	1038.2112	1692.782	Holt's Exponential Smoothing
05Q41	591.200919	6488.1649	153.92435	6143.139	152.9456	ARIMA
05Q43	6910.4297	7026.7198	8137.009	6492.9291	1952.376	ARIMA
05Q44	2455.79438	3198.3604	318.86618	2490.5508	453.2444	Brown's Exponential Smoothing
05Q53	4319.91179	550.8932	6434.4338	548.25312	4743.03	Damp Trend Exponential Smoothing
05Q54	3797.32743	249.19037	4984.0423	257.36922	10474.17	Holt's Exponential Smoothing
05Q61	49.7769478	721.44821	60.553109	729.86046	91.5388	Simple Exponential Smoothing
05QStable	0	0	0	0	0	All the same
05Random	98443.836	67603.255	94553.522	64252.639	60328.78	ARIMA
06M00	1283.0999	632.08717	1596.5618	635.24921	736.7637	Holt's Exponential Smoothing
06M02	727.884687	443.93756	954.03181	449.02756	456.4454	Holt's Exponential Smoothing
06M03	366.5206	348.21087	508.83775	307.77121	142.8623	ARIMA

06M04	611.627348	520.33872	430.11056	614.84909	175.6239	ARIMA
06M11	1777.48642	1599.4357	1326.477	1696.1111	669.7827	ARIMA
06M12	1935.67469	1857.5106	2442.6703	1721.5822	946.4714	ARIMA
06M14	123.743681	183.77456	59.819369	277.31137	66.09994	Brown's Exponential Smoothing
06M20	10732.6537	10110.267	15650.857	9987.614	962.8334	ARIMA
06M21	1476.50385	488.68175	1539.3184	430.28301	1221.348	Damp Trend Exponential Smoothing
06M22	4097.97913	2508.2971	2862.81	2570.2719	251.4946	ARIMA
06M23	335.476072	2094.3995	1743.9445	2133.5952	524.1078	Simple Exponential Smoothing
06M24	506.582035	347.3954	170.00532	469.70561	94.63368	ARIMA
06M30	963.749596	1278.8341	90.370706	1442.1205	72.24413	ARIMA
06M31	757.331236	982.34854	97.863916	1006.4181	73.15144	ARIMA
06M33	2373.86315	2286.5809	639.51479	2406.1725	99.50336	ARIMA
06M34	2223.82104	4141.858	517.668	3737.2129	219.1342	ARIMA
06M50	23486.1135	8509.9417	29100.249	7653.3683	22410.83	Damp Trend Exponential Smoothing
06M51	137.507602	456.61249	1789.3549	449.53446	1704.651	Simple Exponential Smoothing

06M52	652.764933	3005.8384	1268.4727	1451.7756	1353.708	Simple Exponential Smoothing
06M60	1438.96171	1221.4566	503.46251	1858.0226	616.5552	Brown's Exponential Smoothing
06M61	58.3830482	67.001469	4.2640385	68.563048	7.0625	Brown's Exponential Smoothing
06M62	38.6591015	24.814764	59.974831	25.566394	11.72917	ARIMA
06M63	164.686367	168.01769	65.826406	138.7931	85.47578	Brown's Exponential Smoothing
06M70	3556.66717	721.30937	4009.0914	693.72717	2313.606	Damp Trend Exponential Smoothing
06M72	4539.76798	4162.2384	2636.4177	4601.264	1700.025	ARIMA
06M84	5210.37827	8254.1618	2173.3775	7975.7619	1114.417	ARIMA
06M91	1908.17326	366.50452	2243.9034	546.17608	1411.763	Holt's Exponential Smoothing
06M92	998.423932	430.52578	1157.3884	437.42754	944.1761	Holt's Exponential Smoothing
06M94	1154.88396	705.78703	1584.5424	723.55441	877.3385	Holt's Exponential Smoothing
06MW	25892.0476	27856.904	35127.173	22591.336	28035.79	Damp Trend Exponential Smoothing
06Q01	6020.24916	6656.5477	2664.1399	5559.4783	1419.034	ARIMA
06Q02	10039.1408	10828.699	12657.921	9735.5296	2101.755	ARIMA
06Q03	2814.19169	3644.5262	378.18248	2861.6353	533.5014	Brown's Exponential Smoothing

06Q04	1675.99039	9236.9037	916.16011	8691.935	495.173	ARIMA
06Q11	18117.468	9511.1695	4555.5287	9303.9234	626.9357	ARIMA
06Q20	4165.66158	4980.6982	6986.1317	4914.039	340.895	ARIMA
06Q21	194.443735	337.70564	2424.8235	334.82875	59.59484	ARIMA
06Q22	8033.86627	10123.998	6996.1915	10261.763	405.8708	ARIMA
06Q23	120.450111	379.253	2601.9952	381.55184	115.7711	ARIMA
06Q24	280.562318	521.88448	291.40019	512.94136	32.68067	ARIMA
06Q32	9833.59668	2334.4684	14191.377	2501.2443	5112.302	Holt's Exponential Smoothing
06Q33	2632.5009	1254.0835	5520.9917	2006.3951	3123.594	Holt's Exponential Smoothing
06Q40	7843.05374	665.31198	8661.948	1270.1067	4707.316	Holt's Exponential Smoothing
06Q41	185.962993	1237.3878	368.72025	1259.9957	421.0994	Simple Exponential Smoothing
06Q42	3268.82359	760.28063	4091.541	761.83618	1339.117	Holt's Exponential Smoothing
06Q44	2250.08638	2166.6432	2427.7032	2103.2424	724.9982	ARIMA
06Q60	3368.54684	920.85643	2685.1526	888.54685	5144.167	Damp Trend Exponential Smoothing
06Q62	89.6136313	70.836862	1827.0136	68.441634	4065.617	Damp Trend Exponential Smoothing

06Q63	6334.32287	1168.6079	19228.587	1238.3104	18970.08	Holt's Exponential Smoothing
06Q64	8985.78098	671.23079	11413.034	667.23092	6209.593	Damp Trend Exponential Smoothing
06QStable	0	0	0	0	0	All the same
06Random	240320.807	142097.2	237343.49	134934.77	167060.9	Damp Trend Exponential Smoothing

Table A-2: The complete distribution of local optimal solutions based on business simulation reports

References

Aberdeen Group (2010). *State of the Market: Retail Survival Strategies for 2008, 2009.* Retrieved from http://www.aberdeen.com/

Aviv, Y. (2003). A time series framework for supply-chain inventory management. *Operations Research*, Vol. 51, No. 2, pp. 210-227.

Billah, B., King, M. L., Snyder, R. D., & Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International journal of forecasting*, Vol. 22, pp. 239– 247.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. New York, NY: Oxford University Press.

Bodily, S. E., & Freeland, J. R. (1988). A simulation of techniques for forecasting shipments using firm orders-to-date. *The journal of the operational research society*, Vol. 39, No. 9, pp. 833-846.

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for K-Means clustering. *Microsoft research report.* MSR-TR-98-36.

Bradley, P. S., Fayyad, U. M., & Reina, C. A. (1998). Scaling EM (Expectation-Maximization) clustering to large databases. *Microsoft research report.* MSR-TR-98-35.

Brown, R. G. (1959). *Statistical forecasting for inventory control.* New York, NY: McGraw Hill.

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis of example* (4th ed.). Malden MA: Wiley-Interscience.

Chen, F., Ryan, J. K., & Simchi-Levi3, D. (2000). The impact of exponential smoothing forecasts on the bullwhip effect. *Naval research logistics*, Vol. 47, pp. 269-286.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE transaction on knowledge and data engineering*, Vol. 8, No. 6, pp. 866-883.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM,* Vol. 13, No. 6, pp. 377-387.

Connor, J. T., Martin, R. D., & Atlas, L. E. (1994). Recurrent neural networks and robust time series prediction. *IEEE Transactions on neural networks*, Vol. 5, No. 2, pp. 240-254.

Contreras, J., Espínola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next day electricity prices. *IEEE transactions on power systems*, Vol. 18, No. 3, pp. 1014-1021.

Dempster, A. P., Laird, N. M., & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal statistical Society,* Series B, Vol. 39, No. 1, pp.1-38,

Eisen, M. B. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the national academic science*, Vol. 95, pp. 14863-14868.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: The MIT Press.

Fildes, R. (1979). Quantitative forecasting – the state of the art: extrapolative models. *Journal of the operational research society*, Vol. 30, No. 8, pp. 691-710.

Frank, R. J., Davey, N., & Hunt, S. P. (2001). Time series prediction and neural networks. *Journal of intelligent and robotic systems*. Vol. 31, No. 1-3/May, pp. 91-103.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15, pp. 246-263.

Gardner, E. S. (1985). Exponential smoothing: the state of the art. *Journal of forecasting*, Vol. 4, No. 1, pp. 1– 28.

Gardner, E. S. (1990). Evaluating forecast performance in an inventory control system. *Management science*, Vol. 36, No. 4, pp. 490-499.

Gardner, E.S., McKenzie, J.R., & McKenzie, E.D. (1989). Seasonal exponential smoothing with damped trends. *Management science*, Vol. 35, No. 3, pp. 372-376.

Gershenfeld, N. A., & Weigend, A. S. (1993). *The future of time series*. Boston, MA: Addison-Wesley.

Geurts, M. D., & Ibrahim, I. B. (1975). Comparing the Box-Jenkins approach with the exponentially smoothed forecasting model application to Hawaii tourists. *Journal of marketing research*, Vol. 12, No. 2, pp. 182-188.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Harrison, P. J. (1967). Exponential smoothing and short-term sales forecasting. *Management science*, Vol. 13, No. 11, pp. 821-842.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural networks for perception*,Vol. 2, pp. 65-93.

Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: an evaluation. *Management science*, Vol. 27, No. 2, pp. 115-138.

Holt, C. C. (1957). Forecasting trends and seasonal by exponentially weighted averages. *ONR Memorandum*, No. 52. Reprinted in 2004 on *International journal of forecasting*, Vol. 20, No. 1, pp. 5–13.

Industry Canada and Retail Council of Canada (2010). *State of retail: The Canadian report 2010*. Retrieved from www.ic.gc.ca/retail.

Jain, A. K., & Dubes, R. C. (1998). *Algorithms for clustering data*. Upper Saddle River, NJ: Prentice-Hall.

Johnson, G., & Thompson, H. (1975). Optimality of myopic inventory policies for certain dependent demand processes. *Management Science*, Vol. 21, pp. 1303-1307.

Kantardzic, M. (2003). *Data mining: concepts, models, methods, and algorithms*. Malden MA: Wiley-Interscience.

Keogh, E., & Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, Vol. 8, No. 2, pp. 154-177.

Klosgen,W., & Zytkow, J. M. (2002). *Handbook of data mining and knowledge discovery*. New York, NY: Oxford University Press Inc..

Kohonen, T. (1979). Self-organzing maps. *Springer series in information sciences*, Vol. 30, pp. 426.

Kruger, G. A. (2005). A statistician looks at inventory management. *Quality progress*, Vol. 38 No. 2, pp. 36.

Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. Malden MA: Wiley-interscience.

Lawrence, J. (1993). *Introduction to neural networks: design, theory, and applications*. Nevada City, CA: California Scientific Software Press.

Lee, T. S., & Adam, Jr, Everett E. (1986). Forecasting error evaluation in Material Requirements Planning (MRP) production-inventory systems. *Management science*, Vol. 32, No. 9, pp. 1186-1205.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, Vol. 38, No. 11, pp. 1857-1874.

Lim, C., & McAleer, M. (2001). Forecasting tourist arrivals. *Annals of tourism research*, Vol. 28, No. 4, pp. 965-977.

Lingras, P., & Akerkar, R. (2008). *Building an intelligent web: theory and practice*. Sudbury, MA: Jones and Bartlett Learning.

Lingras, P., Zhong, M., & Sharma, S. (2008). Evolutionary regression and neural imputations of missing values. B. Prasad (Ed.): *Soft Computing applications in industry, STUDFUZZ*, Vol. 226, pp. 151–163.

Mendenhall, W., & Sincich, T. (1995). *Statistics for engineering and science*. Upper Saddle River, NJ: Prentice-Hall.

Pearson, K., Yule, G. U., Blanchard, N., & Lee, A. (1903). The law of ancestral heredity. *Biometrika*, Vol. 2, No.2, pp. 211-236.

Peterson, R., & Silver, E. A. (1979). *Decision systems for inventory management and production planning*. New York, NY: John Wiley & Sons.

Ray, W. D. (1982). ARIMA forecasting models in inventory control. *The journal of the operational research society*, Vol. 33, No. 6, pp. 567-574.

Silver, E. A. (1981). Operations research in inventory management: a review and critique. *Operations research*, Vol. 29, No. 4, pp. 628-645.

Silver, E. A., Pyke, D. F., & Peterson, R. (1998). *Inventory management and production planning and scheduling*. New York, NY: Wiley.

Snyder, R. D., Koehler, A. B., & Keith Ord, J. (2002). Forecasting for inventory control with exponential smoothing. *International journal of forecasting*, Vol. 18, No. 1, pp. 5-18.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., & Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing*, Vol.70, No. 16-18, pp. 2861-2869.

Tamayo, P. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the national academic science*, Vol. 16, No. 96, pp. 2907-2912.

Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *The journal of the operational research society*, Vol. 54, No. 8, pp. 799-805.

Tseng, F. M., Yu, H. C., & Tzeng, G. H. (2002). Combining neural network model with seasonal time series ARIMA model. *Technological forecasting & social change*, Vol. 69, pp. 71–87.

Van Der Voort, M., Dougherty, M., & Watson. S. (1996). Combining kohonen maps with arima time series models to forecast traffic flow. *Transportation research part C: emerging technologies*, Volume 4, No. 5, pp. 307-318.

Ward, J. H. (1963). Hierarchical Grouping to optimize an objective function. *Journal of American Statistical Association, Vol. 58*, No. 301, pp. 236-244.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, No. 6, pp. 324–342.

Yule, G. U. (1897). On the theory of correlation. *Journal of the Royal Statistical Society*, Vol. 60, No. 4, pp. 812-854.