ARTIFICIAL INTELLIGENCE FOR DATA MINING IN THE CONTEXT OF ENTERPRISE SYSTEMS

Réal Carbonneau

A Thesis
In
The John Molson School of Business

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Administration at
Concordia University
Montréal, Québec, Canada

August 2005

Your file Votre référence
ISBN: 0-494-10321-3
Our file Notre référence
ISBN: 0-494-10321-3

ABSTRACT

Artificial Intelligence for Data Mining in the context of Enterprise Systems

Réal Carbonneau

Effective supply chain management is one of the key determinants of success of today's businesses. A supply chain consists of various participating businesses that ultimately provide value to the end consumer. However, communication patterns that emerge in a supply chain tend to distort the original consumer's demand and create high levels of noise (randomness). This distortion and noise negatively impact forecast quality of the participants. In this thesis we comparatively examine the quality of new artificial intelligence (AI) based forecasting techniques to identify if they can provide increased forecast accuracy in such conditions. Experiments are performed using data from a chocolate manufacturer, a toner cartridge manufacturer, as well as data from the Statistics Canada manufacturing survey.

A representative set of traditional and AI-based forecasting techniques are applied to the demand data and the accuracy of the methods are compared. As a group, the average performance in terms of rank of the AI techniques does not outperform the traditional approaches. However, using a support vector machine (SVM) that is trained on multiple demand series as one general model, which we call a Super Wide model, has produced the most accurate forecast. Providing a large number of examples to this specific AI technique was the key to achieving high quality forecasts. This leads us to conclude that the best AI technique does outperform its best traditional counterpart, which is exponential smoothing.

Table of Contents

List of Figures

List of Tables

1 Introduction

Central to much of today's business is the notion of Supply Chain, where resources are combined to provide value to the end consumer. Recently, firms have begun to realize the importance of integration across the stakeholders in the supply chain. Such integration often relies heavily on, or at very least, includes sharing information between various business partners (Zhao, Xie et al. 2002). Although such initiatives could potentially reduce forecast errors, they are neither ubiquitous nor complete and forecast errors still abound. Collaborative forecasting and replenishment (CFAR) permits a firm and its supplier-firm to coordinate decisions by exchanging complex decision-support models and strategies, thus facilitating integration of forecasting and production schedules (Raghunathan 1999). However, in the absence of CFAR, firms are relegated to traditional forecasting and production scheduling. As a result, the firm's demand (e.g., the manufacturer's demand) appears to fluctuate in a random fashion, even if the final customer's demand has a predictable pattern. Forecasting the manufacturer's demand under these conditions becomes a challenging task due to what is known as "bullwhip effect" (Lee, Padmanabhan et al. 1997a) – a result of information asymmetry.

The objective of this research is to study the feasibility of forecasting the distorted demand signals in the extended supply chain using adaptive artificial intelligence techniques as compared to traditional techniques. More specifically, the work focuses on forecasting the demand at the upstream (manufacturer's) end of the supply chain. The main challenge lies in the distortion of the demand signal as it travels through the extended supply chain. Our use of

the term extended supply chain reflects both the holistic notion of supply chain as presented by Tan (2001) and the idealistic collaborative relationships suggested by Davis and Spekman (2004).

The value of information sharing across the supply chain is widely recognized as the means of combating demand signal distortion (Lee, Padmanabhan et al. 1997b). However, there is a gap between the ideal of integrated supply chains and the reality (Gunasekaran and Ngai 2004). Researchers have identified several factors that could hinder such long-term stable collaborative efforts. The subsequent paragraphs will discuss a broad insight on supply chain collaboration issues.

Premkumar (2000) lists some critical issues that must be addressed in order to permit successful supply chain collaboration, including:

- alignment of business interests;
- long-term relationship management;
- reluctance to share information;
- complexity of large-scale supply chain management;
- competence of personnel supporting supply chain management;
- performance measurement and incentive systems to support supply chain management.

In most companies these issues have not yet been addressed in any attempts to enable effective extended supply chain collaboration (Davis and Spekman 2004). Moreover, in many supply chains there are power regimes and power sub-regimes that can prevent supply chain optimization (Cox, Sanderson et al. 2001), "Hence, even if it is technically feasible to integrate

systems and share information, organizationally it may not be feasible because it may cause major upheavals in the power structure" (Premkumar 2000). Furthermore, it has been mathematically demonstrated, that while participants in supply chains may gain certain benefits from advance information sharing, it actually tends to increase the bullwhip effect (Thonemann 2002).

Another complicating factor is the possibility of the introduction of inaccurate information into the system. Inaccurate information too would lead to demand distortion, as was reported in a study of the telecom industry demand chain. In the study, some partners were double forecasting and ration gaming (Heikkila 2002), ordering more than they need, despite the fact that there was a collaborative system in place and a push for the correct use of this system.

Finally, with the advance of E-business there is an increasing tendency towards more "dynamic" (Vakharia 2002) and "agile" (Gunasekaran and Ngai 2004; Yusuf, Gunasekaran et al. 2004) supply chains. While this trend enables the supply chains to be more flexible and adaptive, it could discourage companies from investing into forming long-term collaborative relationships among each other due to the restrictive nature of such commitments. For example, collaborative supply chain strategies may involve exchanging sensitive information, which leads to significant risks, or connecting information systems between two partners, which results in large costs. Because such initiatives put major emphasis in long-term commitments, they could be in conflict with business objectives of high agility and adaptability. Depending on specific situations, some organizations may forgo some collaboration and integration initiatives to retain agility and flexibility.

The above reasons are likely to impede extended supply chain collaboration, or may cause inaccurate demand forecasts in information sharing systems. In addition, the current realities of businesses are that most extended supply chains are not collaborative all the way upstream to the manufacturer and beyond, and, in practice, are nothing more than a series of dyadic relationships (Davis and Spekman 2004). In light of the above considerations, the problem of forecasting distorted demand is of significant importance to businesses, especially those operating towards the upstream end of the extended supply chain.

In the present investigation of the feasibility and comparative analysis of artificial intelligence approaches to forecasting manufacturer's distorted demand, we will use advanced tools, including Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Support Vector Machines (SVM). From the most general perspective, these learning techniques permit machines to identify patterns in data. Neural Networks represent an attempt to reproduce the function of a biological brain. This is done by modeling neurons (also called processing elements) and connections among them (which provide a similar function as synapses). The neural network learns by adjusting the weights of the connections between neurons. Recurrent Neural Networks (RNN) are an extension of Artificial Neural Networks to include recurrent connection which permit learning to retain past information. Support Vector Machines are a more recent development, which represent an alternative way to learn patterns in data. Instead of being modeled on a biological analog they have been developed from mathematical theory. The Support Vector Machine learns by identifying support vectors in higher dimensional feature spaces in a way, which attempts to reduce the structural risk.

The performance of these machine-learning methods will be compared against baseline traditional approaches such as exponential smoothing, moving average, linear regression and a model shown in recent research to have very good overall performance, namely the Theta model. The data used for this research comes from three different sources. The first two are from the enterprise systems of a chocolate manufacturer and a toner cartridge manufacturer, who, by the nature of their position in the supply chain, are subject to considerable demand distortion. The third source of data comes from the Statistics Canada manufacturing survey. Inclusion of this survey in the study has the aim of increasing the validity and providing the possibility of replication of results, since the survey data is publicly accessible, while the first two data sources are private.

The remainder of this research report reviews the background and related work for demand forecasting in the extended supply chain; introduces the artificial intelligence approaches included in our analysis; reviews the data sources; presents the research methodology; the experiment implementation and describes and analyses the results of our experiments. The work concludes with the discussion of findings and directions for future research.

2 Background

In the following sections we will review the relevant past work in the areas of supply chain management, traditional forecasting techniques and artificial intelligence- (AI) based forecasting techniques.

2.1 Supply Chain

One of the major purposes of supply chain collaboration is to improve the accuracy of forecasts (Raghunathan 1999). However, since, as discussed above, it is not always possible to have the members of a supply chain work in full collaboration as a team, it is important to study the feasibility of forecasting the distorted demand signal in the extended supply chain in the absence of information from other partners. Incidentally, although minimizing the entire extended supply chain's costs is not the primary focus of this research, it has been demonstrated that improved quality of forecasts will ultimately lead to overall cost savings (Stitt 2004).

In respect to managing the supply chain with AI–based approaches, the use of simulation techniques has shown that genetic algorithm based artificial agents can achieve lower costs than human players. Such agents can even minimize costs lower than the "1-1" policy without explicit information sharing (Kimbrough, Wu et al. 2002). The "1-1" policy means that the supply chain member orders exactly the same amount that was ordered from them. However, the above study was based on global optimization of the problem and is only useful in situations where there is central planning for the complete supply chain.

Analysis of forecasting techniques is of considerable value for firms, as it has been shown that the use of moving average, naïve forecasting or demand signal processing will induce the bullwhip effect (Dejonckheere, Disney et al. 2003). Autoregressive linear forecasting, on the other hand, has been shown to diminish bullwhip effects, while outperforming naïve and exponential smoothing methods (Chandra and Grabis 2005). In this research, we will analyze the applicability of artificial intelligence techniques to demand forecasting in supply chains.

The primary focus of this work is on facilitating demand forecasting by the members at the upstream end of a supply chain. The source of the demand distortion in the extended supply chain simulation is demand signal processing by all members in the supply chain (Forrester 1961). According to Lee, Padmanabhan et al. (1997b), demand signal processing means that each party in the supply chain does some processing on the demand signal, thus transforming it before passing it along to the next member. As the end-customer's demand signal moves up the supply chain, it is increasingly distorted because of demand signal processing. This occurs even if the demand signal processing function is identical in all parties of the extended supply chain. For example, even if all supply chain members use a 6 month trend to forecast demand, distortion will still occur. The phenomenon could be explained in terms of the chaos theory, where a small variation in the input could result in large, seemingly random behavior in the output of the chaotic system. This is demonstrated graphically on Figure 1.

The top portion of Figure 1 depicts a model of the extended supply chain with increasing demand distortion, which includes a collaboration barrier. The barrier could be defined as the link in the supply chain at which no explicit forecast information sharing occurs between the partners. In the bottom portion of Figure 1 we depict the effect of distorted demand

forecasting. Using the manufacturer's historical information on the distorted demand, future demand is forecasted. Between each link in the chain, distortion occurs. Thus, our objective is to forecast future demand (far enough in advance to compensate for the manufacturer's lead time) based only on past manufacturer's orders. To this end, we will investigate the utility of adaptive AI techniques. Consequently, if an increase in forecasting accuracy can be achieved, it will result in lower costs because of reduced inventory as well as increased customer satisfaction that will result from an increase in on-time deliveries (Stitt 2004).

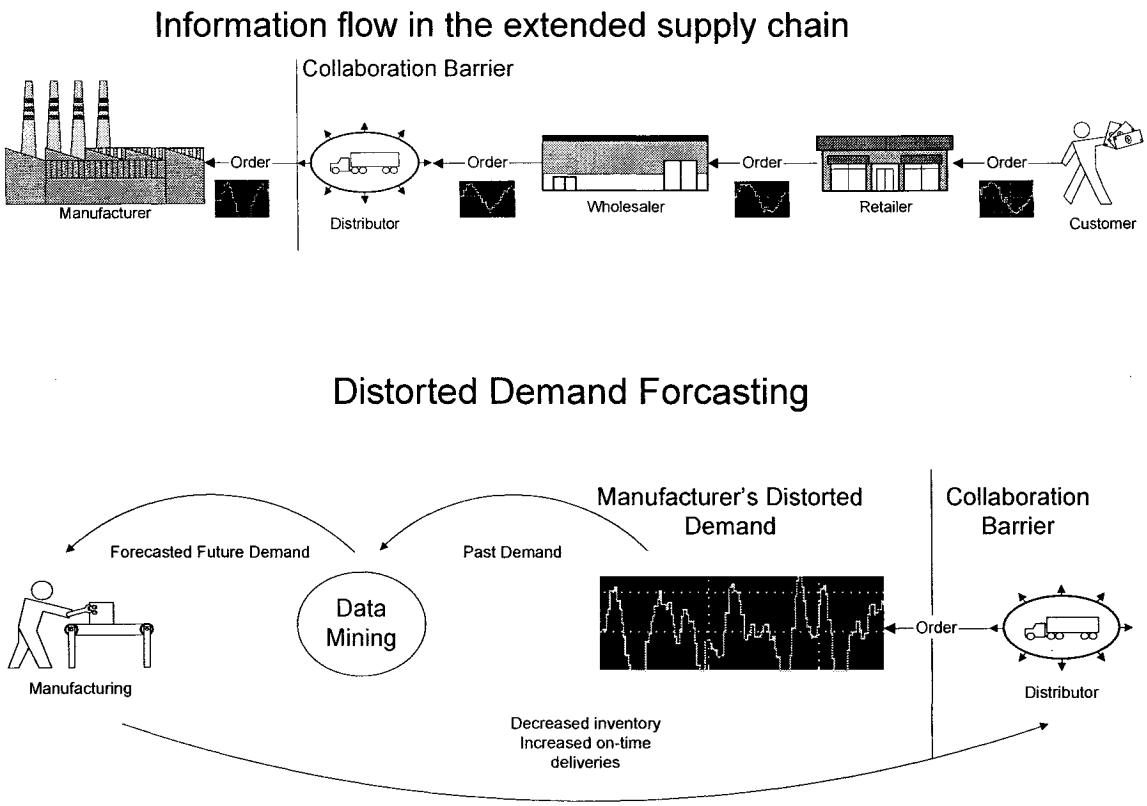

Figure 1 - Distorted demand signal in an extended supply chain

2.2 Traditional Forecasting

Since a fundamental part of efficient management of the supply chain relies on forecasting, we will go through an in-depth examination of the current state of forecasting in general and more specifically business and supply chain forecasting. Issues of interest that will be examined are the realities of forecasting faced by practitioners in today's businesses, the techniques and software commonly used, as well as academic research in the field.

2.2.1 *Practitioners' Forecasting Overview*

In this section we present the results of studies on the existing methods used by practitioners for business forecasting. Jain (2004c) notes that in general there are three types of models commonly used in the industry, including: Time-series, Cause-and-Effect and Judgmental. Time-series forecasting is by far the most commonly used approach (71%) because of its simplicity and the availability of data, followed by Cause-and-Effect models (19%) and Judgmental models (10%). In the category of Time-Series forecasting there are many specific models, but the most common ones are by far the simplest ones with Averages and Simple Trend used 65% of the time, next is Exponential-smoothing at 25% and then Box-Jenkins (6%) and Decomposition methods (5%). The Cause-and-Effect modeling is most often executed with Regression analysis (74%), then Econometric models (21%), and Neural Networks (5%) (Jain 2004c).

Furthermore, Jain (2004b) indicates that when these models are implemented using computer software; 63% of the forecasting market share is held by Excel; 36% by various forecasting systems (both stand-alone and integrated); and 1% with Lotus 1-2-3. We can see a clear dominance by the spreadsheet forecasting tools, which also translates into requirement for a certain amount of human intervention for executing most forecasting tasks. Of the stand-alone forecasting software, John Gault has a large part of the market (46%) followed by SAS (28%). Forecast Pro has an 11% market share and there are a few other products that have smaller presence as represented in Figure 2. Of the integrated software solutions, the market share is as follows; SAP (23%), Manugistics (16%), Demand Solutions (13%), I2 Technology (11%), Oracle (11%) and others follow at 4% and under (Figure 3) (Jain 2004b). From these

numbers we see that most forecasting is still done with time-series analysis in spreadsheet, but that there is a diverse set of integrated solutions that include forecasting solutions which are in use in a minority of businesses.

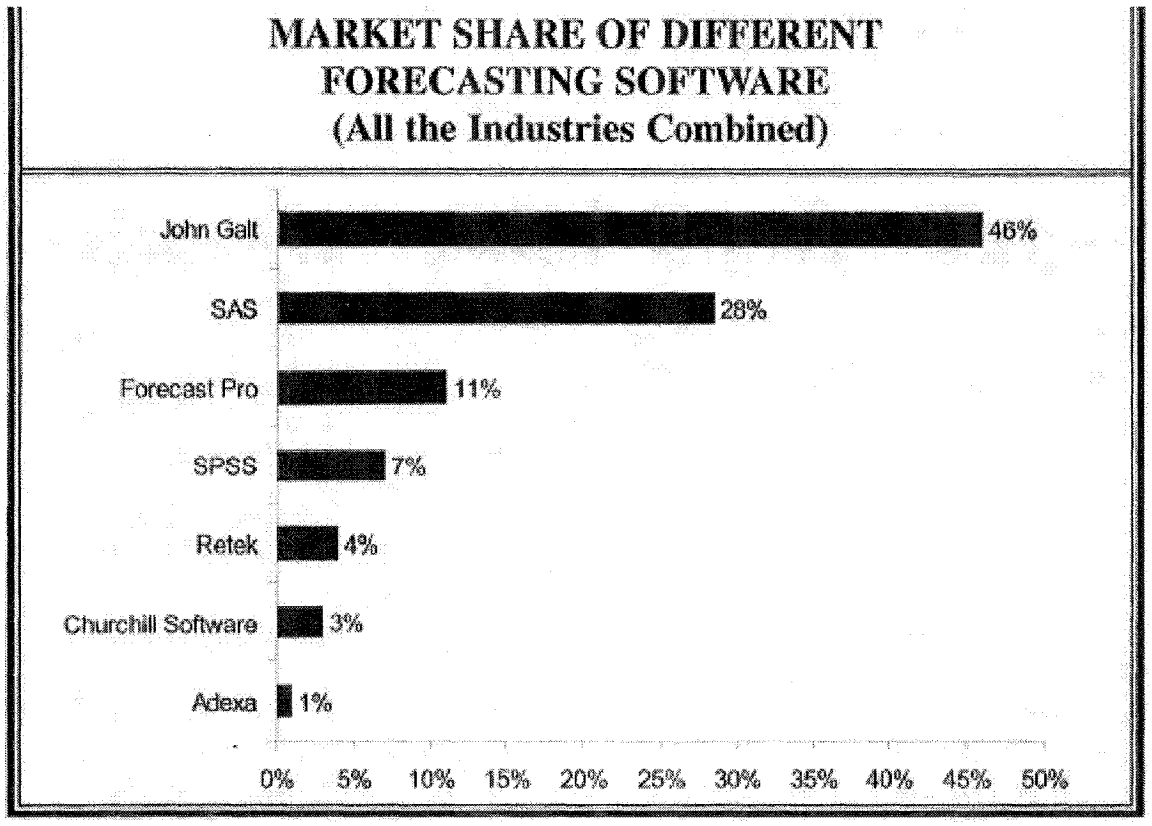

Figure 2 - Market Share of different forecasting software (Jain 2004b)

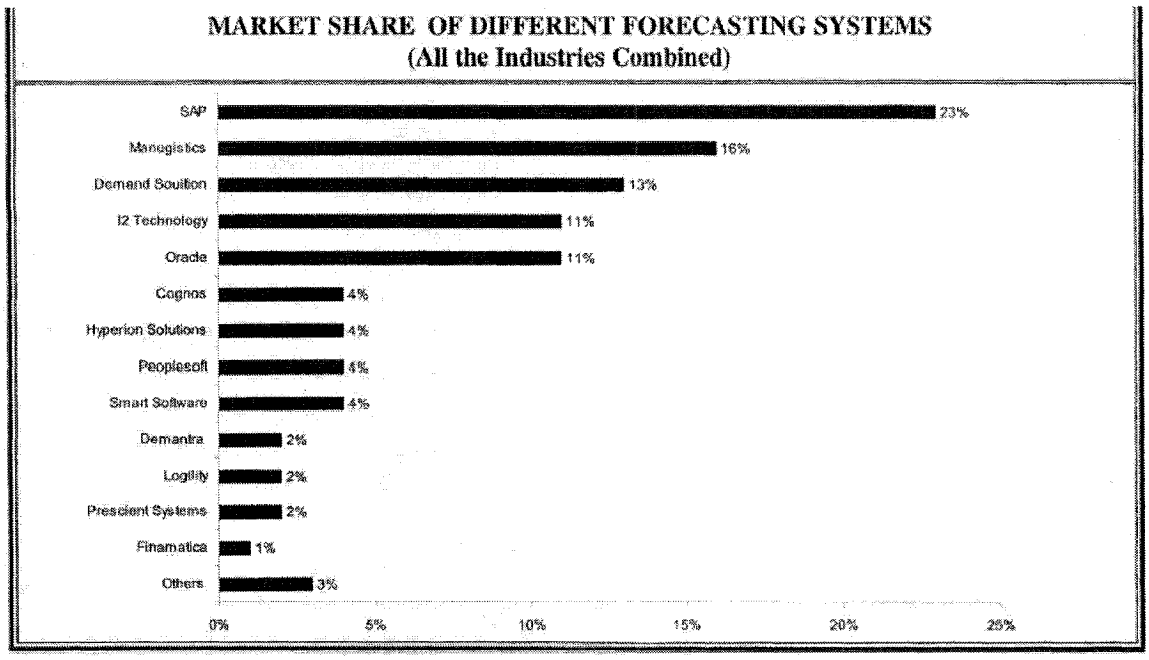

Figure 3 - Market share of different forecasting systems (Jain 2004b)

Any forecast can eventually be compared with actual results at which point a forecast error can be calculated. The forecast error is very important since it is how forecast performance is evaluated and it also permits comparison of forecasting accuracy with other methods and other businesses. The average forecasting error across all industries for 1 month ahead is 26%, for 2 months and 1 quarter ahead is 29% and for 1 year ahead is 30%. However, in every industry, there are some members that have much more accurate forecasts than others, such as in the consumer products industry, where the best in class forecaster has a 10% error for 1 month forecasts and a 14% error for 1 year ahead as compared to the industry average of 26% error for 1 month forecast and 30% error for 1 year ahead forecasts (Jain 2004a). This indicates that there can be significant competitive advantages gained from improved forecast accuracy.

Although the above information might be interesting and important, it should be noted that these details are general across all business-related forecasting functions. Although they do not give us specific information about forecasting for members at the end of the supply chain who usually experience the highest level of demand distortion (which can be highly chaotic and unexplained), they are still important for understanding average forecasting in business, which may often be reflected in specific business functions.

Specific research related to forecasting within the purchasing function of business has been performed by Wisner and Stanley (1994) on 148 respondents to a purchasing forecasting survey. Interestingly, 39% of respondents indicated that they rarely or never used forecasts, which may indicate a lack of data available for forecasting, a lack of simple but effective forecasting techniques, or a lack of skills and resources related to forecasting. Additionally, 73 percent of respondents indicated that they used purchasing forecasting for less then 10 years. With respect to actual forecasting, 67% indicated that they generated their forecasts manually and about half attempted to change the forecast parameters to increase forecast accuracy. Although this adjustment may lead to better forecasts, it is important to note that tuning forecasts may result in high accuracy on the known existing data set while decreasing the accuracy of future forecasts. Therefore, it is interesting that tuning is performed by some users despite the potential danger if not done correctly. A potential overfitting of the forecasting to past data will decrease forecasting accuracy which is not only detrimental to the individual supply chain member, but also to other members since the increased forecasting error drastically increase the overall supply chain demand distortion in a cascading fashion.

However, this might be correctly controlled by 41% of respondents who track forecast error every period.

The top quantitative forecasting techniques are simple moving average (62%), weighted moving average (46%), exponential smoothing (45%) exponential smoothing using trend and seasonal enhancements (39%), simulation model (22%), regression model (21%), econometric model (19%) and Box-Jenkins model (14%) (Wisner and Stanley 1994). It seems that, on average, purchasing managers rely on manual and simple forecasting techniques and that there is a lack of simple but powerful forecasting techniques that can be effectively adopted and relied on by practitioners.

2.2.2 *Forecasting Software Overview*

Since most forecasting is done in Excel (Jain 2004b), it is important to review what forecasting functionalities are available in this software. Excel is a very flexible tool that provides many predefined mathematical and statistical functions, which can be used for forecasting. Additionally, necessary formulas can also be defined by the user. Excel add-in functionalities can be used for more advanced statistical functions such as regression as well as for third-party solutions which can provide a wide range of functionalities ranging from statistical models to neural networks (NeuroSolutions 2005). Although the functionality provided by Excel is extensive, we will consider only the standard functionalities commonly used by practitioners. The simple moving average and trend are available as standard Excel functions, which permit quick and easy application to business problems. Simplicity and ease of use of these techniques might explain their wide acceptance. Exponential smoothing and regression are

available from the standard Excel Statistics Analysis ToolPack (Microsoft 2005) and other forecasting approaches may be developed by the user in Excel or may be purchased as third party add-in solutions.

Specialized forecasting software provide more extensive forecasting functionalities, for example John Galt ForecastX Wizard 6.0 is an Excel Add-In that provides over 30 different forecasting techniques (John Galt Solutions 2005), some examples of which are presented in Table 1.

Table 1 - Example forecasting techniques available in John Galt ForecastX

• Seasonal Models • Box Jenkins (ARIMA modeling) • Census II X-11 (Additive and Multiplicative) • Decomposition (Additive and Multiplicative) • Holt's Winters Exponential Smoothing (Additive and Multiplicative) • Simple Methods (Time Series) • Adaptive Exponential Smoothing • Brown's Double Exponential Smoothing	• Brown's Triple Exponential Smoothing • Exponential Smoothing • Holt's Double Exponential Smoothing • Moving Average • Weighted Moving Average • Casual Forecasting • Linear Regression • Polynomial Regression • Simple Regression • StepWise Regression with Dynamic Lagging

Other competing products, such as SAS, provide the same range of forecasting functionalities (SAS Institute 2005).

Although there is clearly a wide range of forecasting techniques and tools available we do not see such a variety in use by practitioners. Having so many different techniques may actually make the forecasting situation difficult for businesses since it is not clear which one to use in what circumstances. In addition, the users are also supposed to correctly set the parameters for these techniques. How is the time-strapped and information-overloaded practitioner supposed to know what technique to use and to be properly informed about the correct use of a selected technique? Most practitioners do not have the skill or time to do this, thus they continue to use simple techniques or hire an expert to implement such solutions for their business. Ideally, a forecasting solution must represent a turnkey solution to enable higher adoption by practitioners. In this thesis will attempt to identify such a solution.

2.2.3 Forecasting System Overview

Some integrated enterprise solutions also include an integrated forecasting system within the core system. For example, the SAP R/3 system provides integrated forecasting for various scenarios including product demand. The list of standard forecasting models available with the core system is presented below and additional functionalities permit automatic model selection by choosing the model and parameters that provide the lowest error (SAP 2005).

- Exponential smoothing model
- Trend model
- Seasonal model
- Seasonal trend model
- Moving average
- Weighted moving average

- 2nd order exponential smoothing

Although once again we see a wide variety of forecasting techniques available, there is however means to automatically select models and model parameters based on lowest forecast error. Since these forecasting capacities are directly integrated into the enterprise system, they have access to historical data and can be automatically processed via the model selection facility. However, this approach does have a disadvantage in that it can overfit the data by identifying a model that reduces the error for the known data but at the same time increase the error on future unknown data. These automatic model selection procedures are not able to identify the model that will provide the best generalization.

2.2.4 Academic Forecasting Research Overview

There has been extensive research performed in the area of forecasting which has provided a large number of forecasting techniques and algorithms as can be seen by the very existence of specialized journals such as the "Journal of Forecasting" and the "International Journal of Forecasting". There is also forecasting research presented in the many mathematics, statistics, operations management and supply chain journals to name a few categories. This is also evident from the large selection of algorithms in existing forecasting software, which represents only the most successful subset of past forecasting research as seen in the previous section.

Past forecasting competitions have been created with the purpose of identifying the most accurate forecasting methods. Although these projects are named competitions because they

frame the research as a competition between forecasting algorithms for the purpose of identifying the best algorithm, they are indeed research endeavors of which details are published.

There is an extensive series of competitions that have grown in complexity and validity over the years as both academics' and practitioners' critiques have been considered in improving the competitions. One of the first large studies of various forecasting techniques on 111 real micro and macro time-series from a wide range of businesses and industries found that the simpler forecasting methods had better overall accuracy than more complex ones (Makridakis and Hibon 1979). The most formalized versions of theses competitions started with the original M-Competition. This research tested 15 different forecasting techniques over 1001 time series. The 15 forecasting techniques were obtained from a published call to participate, thus attempting to ensure broad representation of the current academic and commercial forecasting techniques. The results lead to the same general conclusions that simpler forecasting techniques performed better overall than more complex ones (Makridakis, Andersen et al. 1982). The reason for this is intuitive: when researchers develop more complex forecasting models they focus on a specific industry, problem domain or approach which permits increased performance of the forecast by specifically tuning the forecasting approach. This increase in accuracy in a specialized subset of real world forecasting is often done unknowingly at the expense of more generalized performance.

The M2-Competition was a variation that occurred in real-time where the participants had to forecast 15 months into the real future and then the performance was determined as time progressed. The results again were similar, with the sophisticated techniques performing the

same or worse than the simple ones (Makridakis, Chatfield et al. 1993). All of these results were extensively verified and reproduced over the years (Makridakis and Hibon 2000).

The subsequent M3 Competition increased the number of time-series to 3003 and selected the time-series from a wide variety real life situations (micro, macro, industry, time interval, number of observations, etc) to provide an extremely wide diversity of forecasting problems. The distribution of the experiments time-series is presented across two dimensions, the time interval and the type of data in Figure 4. Here we see that there are 474 monthly microeconomic series and 334 monthly industry series. The time-series diversity represented in this research helps increase the generalization validity of the results

The classification of the 3003 time series used in the M3-Competition

Time interval between successive observations	Types of time series data						
	Micro	Industry	Macro	Finance	Demographic	Other	Total
Yearly	146	102	83	58	245	11	645
Quarterly	204	83	336	76	57		756
Monthly	474	334	312	145	111	52	1428
Other	4			29		141	174
Total	828	519	731	308	413	204	3003

Figure 4 - M3 timeseries classification (Makridakis and Hibon 2000)

This wide variety of time series was forecasted by 24 different methods ranging from very simple, such as deseasonalized naïve to highly sophisticated such as neural networks as presented in Figure 5. The M3-Competition included academic forecasting methods as well as commercial ones (indicated by an asterisks *). The major categories included in their experiment are Naïve/simple, Explicit trend models, Decomposition, variations of the general ARMA (Auto-Regressive Moving Average) model, expert systems and neural networks. This

is another indication of the wide scope and validity of the M3-Compeition research as a result of the diversity of methods included.

The 24 methods included in the M3-Competition classified into six categories

Method	Competitors	Description
Naïve/simple		
1. Naïve2	M. Hibon	Deseasonalized Naïve (Random Walk)
2. Single	M. Hibon	Single Exponential Smoothing
Explicit trend models		
3. Holt	M. Hibon	Automatic Holt's Linear Exponential Smoothing (two parameter model)
4. Robust-Trend	N. Meade	Non-parametric version of Holt's linear model with median based estimate of trend
5. Winter	M. Hibon	Holt–Winter's linear and seasonal exponential smoothing (two or three parameter model)
6. Dampen	M. Hibon	Dampen Trend Exponential Smoothing
7. PP-autocast[a]	H. Levenbach	Damped Trend Exponential Smoothing
8. Theta-sm	V. Assimakopoulos	Successive smoothing plus a set of rules for dampening the trend
9. Comb S-H-D	M. Hibon	Combining three methods: Single/Holt/Dampen
Decomposition		
10. Theta	V. Assimakopoulos	Specific decomposition technique, projection and combination of the individual components
ARIMA/ARARMA model		
11. B–J automatic	M. Hibon	Box–Jenkins methodology of 'Business Forecast System'
12. Autobox1[a]	D. Reilly	Robust ARIMA univariate Box–Jenkins
13. Autobox2[a]		with/without Intervention Detection
14. Autobox3[a]		
15. AAM1	G. Melard,	Automatic ARIMA modelling with/without
16. AAM2	J.M. Pasteels	intervention analysis
17. ARARMA	N. Meade	Automated Parzen's methodology with Auto regressive filter
Expert system		
18. ForecastPro[a]	R. Goodrich, E. Stellwagen	Selects from among several methods: Exponential Smoothing/Box Jenkins/Poisson and negative binomial models/Croston's Method/Simple Moving Average
19. SmartFcs[a]	C. Smart	Automatic Forecasting Expert System which conducts a forecasting tournament among four exponential smoothing and two moving average methods
20. RBF	M. Adya, S. Armstrong, F. Collopy, M. Kennedy	Rule-based forecasting: using three methods — random walk, linear regression and Holt's, to estimate level and trend, involving corrections, simplification, automatic feature identification and re-calibration
21. Flores/Pearce1	B. Flores,	Expert system that chooses among four methods
22. Flores/Pearce2	S. Pearce	based on the characteristics of the data
23. ForecastX[a]	J. Galt	Runs tests for seasonality and outliers and selects from among several methods: Exponential Smoothing, Box–Jenkins and Croston's method
Neural networks		
24. Automat ANN	K. Ord, S. Balkin	Automated Artificial Neural Networks for forecasting purposes

[a] Commercially available forecasting packages. Professionals employed by those companies generated the forecasts utilized in this Competition.

Figure 5 - M3-Competition methods (Makridakis and Hibon 2000)

The results have confirmed the original conclusions of M-Competition using a new and much enlarged set of data. In addition, it demonstrated, once more, that simple methods developed by practicing forecasters [...] do as well, or in many cases better, than statistically sophisticated ones [...]." (Makridakis and Hibon 2000). There was a new technique that did provide surprisingly good overall performance in the competition:

> *"a new method, Theta, seems to perform extremely well. Although this method seems simple to use [...] and is not based on strong statistical theory, it performs remarkably well across different types of series, forecasting horizons and accuracy measures." (Makridakis and Hibon 2000).*

Considering this conclusion by these authors who have much experience in forecasting and the related competitions and comparison of techniques, the Theta method (Assimakopoulos and Nikolpoulos 2000) presents a very interesting and potentially useful methods for forecasting distorted supply chain demand.

However, even with the above conclusions, there remains an array of forecasting methods that perform well in certain situations and there is still no recommendation for the best forecasting technique to use. Additionally there are concerns with the results of the competitions, such as the extremely large number of time-series required to forecast to enter this competition which made it infeasible for many researchers to participate, especially in areas such as applied artificial intelligence (Chatfield 2001). As a baseline indicator, the compiled list of 16 commentaries on the M3 Competition (Ord 2001) in just one specific issue of the International Journal of Forecasting, indicates a wide variety of opinions and positions on the

results. This further indicates the disagreement over any conclusion with respects to the best forecasting algorithm and also attests to the complexity of the issue resulting in a situation that distances any hopes of practitioners who just want to know what method to use and how to adjust it's parameters to achieve the best possible forecast.

Considering the large number of new forecasting techniques and specialized variations being constantly presented in forecasting journals as well as in the many statistics and mathematics journals, we will not provide an extensive review of all recent forecasting techniques. In light of the above forecasting competition results and other supporting research which confirms those results, it would seem that simple traditional forecasting techniques outperform more sophisticated and complex ones.

2.3 Artificial Intelligence

As can be seen from the above review of the current state of forecasting, there is a multitude of forecasting algorithms available, which offer a range of settings that can permit tuning of the technique. This thus leads to a general problem of finding the best forecasting technique and the best set of parameters for the chosen technique that results in the most accurate forecast. This process is difficult because the more trial and error executed, the higher the probability of finding a solution that does very well on the given data, but that does not forecast well into the future. Even with the use of a testing set, the same problem can occur if performance on the testing set is evaluated many times. The more choices we have in forecasting techniques and parameters, the more problematic this situation becomes.

The complexity of this problem can be reduced by relying on a class of algorithms that are called "universal approximators", which can, by definition, approximate any function, to an arbitrary accuracy. Using such universal approximators, any required function between past and future data can be learned, thus making exiting forecasting techniques a subset of the functions that the universal approximator can learn. As will be reviewed next, artificial intelligence techniques, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are universal approximators and can be used to learn any function.

However, since forecasting time-series such as those in supply chains involve a data domain, which is highly noisy, we only want to learn true patterns in the data that will be repeated in the future, and we want to ignore the noise. An example of a pattern that can be learnt is a seasonal fluctuation, while noise would be the random fluctuations around this seasonal pattern. The seasonal pattern can be predicted, while the noise cannot. Thus, AI-based techniques have two important features that are useful for noisy supply chain forecasting problems. The first is the ability to learn an arbitrary function, the second is the control of the learning process itself, so that only true patterns that will most likely reoccur in the future are learnt and the noise is ignored.

Of specific interest to our research are Artificial Neural Networks, Recurrent Neural Networks, and Support Vector Machines. Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) are frequently used to predict time series (Dorffner 1996; Landt 1997; Herbrich, Keilbach et al. 1999; Giles, Lawrence et al. 2001). In particular, RNN are included in the analysis because the manufacturer's demand is considered a chaotic time-series. RNN perform back-propagation of error through time that permits the Neural Network to

learn patterns through to an arbitrary depth in the time series. This means that even though a time window of data is provided as the input dimension to the RNN, it can match pattern through time that extends further than the provided current time window, because it has recurrent connections. Support Vector Machines (SVM), a more recent learning algorithm that has been developed from statistical learning theory (Vapnik 1995; Vapnik, Golowich et al. 1997), has a very strong mathematical foundation, and has been previously applied to time series analysis (Mukherjee, Osuna et al. 1997; Rüping and Morik 2003)

2.3.1 Neural Networks

Although artificial neural networks could include a wide variety of types, the most commonly used are feed-forward error back-propagation type neural networks. In these networks, the individual elements ("neurons") are organized into layers in such a way that output signals from the neurons of a given layer are passed to all of the neurons of the next layer. Thus, the flow of neural activations goes in one direction only, layer-by-layer (feed-forward). Errors made by the neural network are then used to adjust all the network weights by moving back through the network (error back-propagation). The smallest number of layers is two, namely the input and output layers. More layers, called hidden layers, could be added between the input and the output layer. The function of the hidden layers is to increase the computational power of the neural nets. Artificial Neural Network (ANN) have been proven to be universal approximators assuming that sufficient hidden layer neurons are provided and assuming that the activation function is bounded and non-constant (de Figueiredo 1980; Hornik 1991)

Neural networks weights are tuned to fulfill a required mapping of inputs to the outputs using training algorithms. The common training algorithm for the feed-forward nets is called "error back-propagation" (Rumelhart, Hinton et al. 1986). This is a supervised type of training, where the desired outputs are provided to the ANN during the course of training along with the inputs. The provided input-output couplings are known as training pairs, and the set of given training pairs is called the training set. The Exclusive-Or (XOR) problem is presented in Table 2 as an example of a training set and its components, the training pairs (input-output couples). The problem is presented in binary (numeric) format, where zero (0) is false and one (1) is true. Thus the logic statement of having x1 or x2 true results in true, but if neither x1 or x2 is true or if both x1 and x2 are true, the result is false.

Table 2 - Training Set example for the XOR problem

	Inputs		Output
	x1	x2	y
Training Pair 1	0	0	0
Training Pair 2	0	1	1
Training Pair 3	1	0	1
Training Pair 4	1	1	0

Thus, using the same approach for much more complicated data-sets such as forecasting noisy supply chain demand, a neural network can learn and generalize from the provided set of past data about the patterns, which may be present in the future.

2.3.2 Recurrent Neural Networks

Recurrent neural networks (RNN) allow output signals of some of their neurons to flow back and serve as inputs for the neurons of the same layer or those of the previous layers. RNN serve as a powerful tool for many complex problems, in particular when time series data is involved. The training method called "back-propagation through time" can be applied to train a RNN on a given training set (Werbos 1990). The Elman network implements back-propagation through time as a two layer backpropagation neural network with a one step delayed feedback from the output of the hidden layer to its input (Elman 1990). Figure 6 shows schematically the structure of RNN for the supply chain demand-forecasting problem. The arrows represent connections within the neural network and more importantly the thicker arrows represent recurrent weights which are distinctive of Recurrent Neural Networks

(RNN). As may be noted in Figure 6, only the output of the hidden neurons is used to serve as the input to the neurons of the same layer.

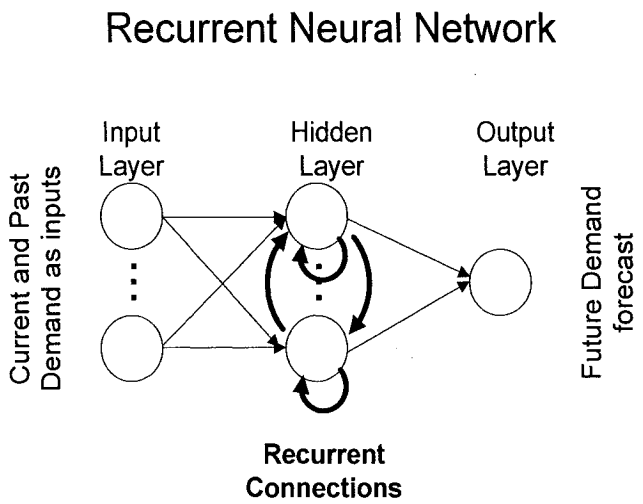

Figure 6 - Recurrent neural network for demand forecasting

2.3.3 Support Vector Machine

Support vector machines (SVM) are a newer type of universal function approximators that are based on the structural risk minimization principle from statistical learning theory (Vapnik 1995) as opposed to the empirical risk minimization principle on which artificial neural networks (ANN) and multiple linear regression (MLR), to name a few, are based. The objective of structural risk minimization is to reduce the true error on an unseen and randomly

selected test example as opposed to ANN and MLR, which minimize the error for the currently seen examples.

Support vector machines project the data into a higher dimensional space and maximize the margins between classes or minimize the error margin for regression using support vectors. Projecting into a higher dimensional space permits identifying patterns that may not be clear in the input space but which become much clearer in a higher dimensional space. Margins are "soft", meaning that a solution can be found even if there are contradicting examples in the training set. The problem is formulated as a convex optimization with no local minima, thus providing a unique solution as opposed to back-propagation neural networks, which may have multiple local minima and, thus cannot guarantee that the global minimum error will be achieved. A complexity parameter permits the adjustment of the level of error versus the model complexity, and different kernels, such as the Radial Basis Function (RBF) kernel, can be used to permit non-linear mapping into the higher dimensional space.

2.4 Research Question

Based on the above review and the conclusion about a lack of clear consensus of which forecasting techniques are best and how these techniques might be used, it would seem that using an advanced artificial intelligence based could provide a simple but powerful solution for forecasting chaotic and noisy distorted customer demand at the end of a supply chain as experienced by manufacturers. This leads us to the question regarding the performances of these techniques: Do artificial intelligence forecasting techniques provide more accurate forecasts of distorted customer demand in a supply chain as experienced by the manufacturer?

In essence, we would like to investigate if AI in general performs better than traditional techniques. Moreover, since it only makes sense for a practitioner to consider and implement only the best available forecasting solution if one can be identified, we will also compare the best representatives from both categories. Thus, we formulate two hypotheses to probe our research question:

H1: Artificial Intelligence based techniques used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better average performance than traditional techniques.

H2: Artificial Intelligence based techniques used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better rank performance than traditional techniques.

H3: The best Artificial Intelligence based technique used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better performance than the best traditional technique.

2.5 Experiment Outline

To answer our research question, we will conduct an experiment to compare the accuracy of artificial intelligence forecasting techniques with traditional forecasting techniques in the context of noisy supply chain demand as seen by the manufacturer. Included in the set traditional forecasting techniques are the simple traditional forecasting techniques, since, as seen previously, these have consistently had overall better performance. Additionally, based on

the results and conclusions of the M3-Competition (Makridakis and Hibon 2000), we include the Theta method (Assimakopoulos and Nikolpoulos 2000) in the group of traditional forecasting techniques. For completeness we also include the classic ARMA (Auto-Regressive Moving Average), sometimes also referred to as the Box-Jenkins model (Box, Jenkins et al. 1994). Even though it was not shown to be a top performer in the previously reviewed research (Makridakis and Hibon 2000), it is included because it is so common as seen by the 7 of the 24 strategies in the M3-Competition are ARMA derivatives as well as two of the expert systems including it (Makridakis and Hibon 2000). The artificial intelligence based forecasting techniques are limited to the 3 general classes presented previously, Artificial Neural Networks (ANN), Artificial Neural Networks (RNN) and Support Vector Machines (SVM). The techniques compared are as follows:

- Moving Average
- Trend
- Exponential Smoothing
- Theta Model (Assimakopoulos & Nikolopoulos 1999)
- Multiple Linear Regression (Auto-Regression)
- Auto-Regressive and Moving Average (ARMA) (Box and al. 1994)
- Neural Networks
- Recurrent Neural Networks
- Support Vector Machines

With the above techniques, we will forecast demand using data for products from three sources.

1. A chocolate manufacturer

2. A toner cartridge manufacturer
3. The Statistics Canada manufacturing survey

2.6 Sample Size and Statistical Power

The following statistical power analysis is based on recommendations from Russel (2001) for determining sample size. Determining the sample size to answer our research question is very difficult, simply because the variance of a forecasting technique on an unknown data set is extremely difficult to estimate in advance. One can vaguely guess that the deviation of a forecast would be the same as the average forecast error. Since an average forecast error of 26% has been identified (Jain 2004a), we will use this as the estimate average deviation. However, since we are concerned with only the manufacturer's end of the supply chain, who by the nature of their position in the supply chain experience extremely noisy demand, this estimate is at the most conservative end of the spectrum.

The next issue to determine is the minimum size of the difference between the means that we want to detect. To determine this, we must identify what level of forecasting accuracy increase is useful and this depends of the potential cost savings. An in-depth survey of six companies has identified the inventory carrying costs to be between 14% and 43% (Lambert and Lalonde 1976). A more recent survey by the Institutes of Management and Administration as reported by the Controller's Report (Anonymous 2005) has also identified most inventory carrying costs to be between 10% and 40% (Figure 7) and represents an average holding cost of around 21%.

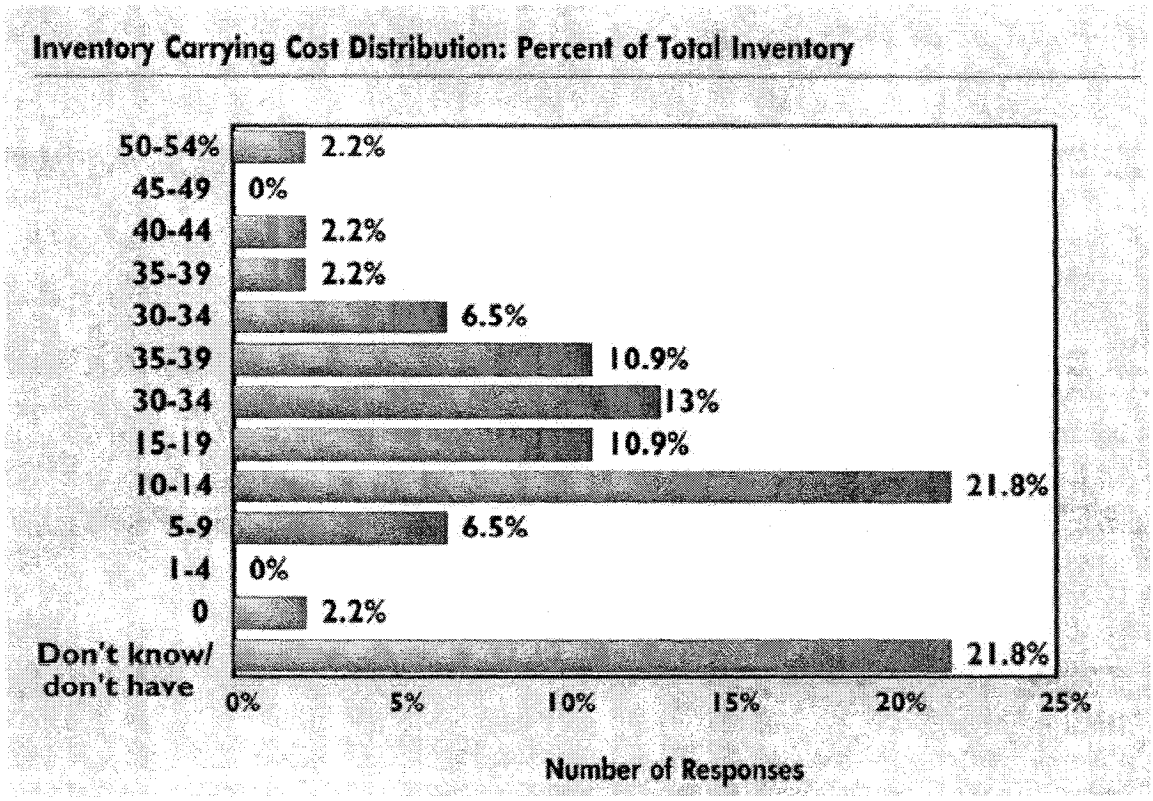

Figure 7 - Distribution of Inventory Carrying Costs (Anonymous 2005)

It is also interesting that about 68% of the respondents to this survey were manufacturers, which means that these inventory carrying costs have a higher relevance for the current research. Using a conservative average inventory carrying cost of 20%, we can make some very rough calculations. For example, a company that has an inventory of 10 million dollars may have a inventory carrying cost of 2 million dollars a year, therefore increasing forecasting accuracy by 1% would lead to a cost saving of about 20,000$ a year. We can also take a large manufacturer as an example, such as Proctor and Gamble that has 4.4 billion dollars of inventory (Proctor and Gamble 2005). From this, we could guess an approximate 880 million dollars of inventory carrying costs, thus, a 1% increase in forecasting accuracy may results in a

cost saving of 8.8 million dollars a year. Clearly these numbers are highly speculative. However, the purpose is that, in the absence of exact figure, we still want to identify the approximate size of forecasting accuracy improvement that we want to detect, and a 1% minimum would seem reasonable.

To detect a 1% change for a dataset that has an estimated minimum standard deviation of 26%, and to do so at a 0.05 significance level for a one-tailed test, would require a minimum of 1833 observations to have a statistical power of 50%. A statistical power of 95% would require 7317 observations. Clearly, the current problem requires very large samples sizes. However, we are still limited by other concerns such as the availability of data and the processing time of all the models required in this experiment. We target a minimum sample size of over 2000 and as will be seen later on, for various reasons, most of the answers will come from a sample size of 2200 observations that come from two problem domain specific datasets. From this sample size and the above parameters, at a 1% change, we have a statistical power of 53%, at a 2% change, we have a statistical power of 96% and at a 3% change we have a statistical power of 99.98%.

2.7 Summary

Based on a review of existing research on supply chains, traditional forecasting techniques and artificial intelligence, we intend to compare traditional and artificial intelligence based forecasting techniques for manufacturers since they experience high levels of distorted demand. A set of traditional forecasting techniques that have been shown to have good performance or are very commonly used is compared with the most common artificial

intelligence forecasting approaches. This question will be answered by doing an experiment with actual data from manufacturers, and this will be done with a data set size of about 2000. A detailed presentation of the data sources and the data is presented in the next chapter

3 Demand Data Definition and Description

3.1 Data Sources

Since we are interested in studying the application of artificial intelligence for forecasting at the manufacturer's end of the supply chain where the original customer's demand will be the most distorted, we require demand data from a manufacturer. However, data records that have been manipulated for various reporting reasons or records that are subject to reporting errors may have a negative impact on the validity of the results of the current research. When manual systems or disparate information systems are used for processing orders, deliveries, inventory, and production, there is a high probability of error or alteration in actual demand values, since there are rarely efficient controls for preventing these.

Thus, accurate data records are important since we are concerned with providing models that can work automatically in the context of an enterprise system. Integrated enterprise systems, such as SAP, are built in such a way that every product or service that goes in and out of the system is controlled as are all monetary transactions. Most transactions such as sales, deliveries and inventory adjustments are not only controlled within their own logistics functions but also integrally controlled by the accounting core since all these business activities affect financial accounts. For these reasons, we only considered manufacturers that are using enterprise systems. There are two manufacturers' who provided demand data as extracted from their ERP systems.

3.1.1 Chocolate Manufacturer

The first manufacturer produces chocolate starting from the cocoa beans, which must be roasted, converted into cacao and then combined with other ingredients according to recipes. This chocolate manufacturer conducts business in both Canada and the United States, therefore the geographic scope of the data is all of North America.

Since the inception of the ERP system and up to the time of the extraction of the data for this research, information regarding demand is available for slightly more than 4 years, from early September 2000 to late September 2004. However, we will not use data from the first month since it is from a system-change over period and may have missing or inaccurate data and we will not use data from the last month since the data was extracted during that month and is incomplete. Therefore, the effective range of our usable data is from October 1, 2000 to August 31, 2004, which represents 3 years and 11 months of data or 47 periods. The data extracted from the enterprise system represents the demand for a product by month.

Because of the number of forecasting models that must be processed and the complexity of some of the models, the number of time-series that are retained for the experiment must be limited. We will be considering the top 100 products with the most cumulative volume during this period since these are the most important products for the manufacturer and are the highly relevant for accurate forecasting. Even though these top 100 products represent only 6% of the products sold by this manufacturer, they account for over 34% of the total sales volume by weight for this manufacturer; thus, it is a highly representative sample of product demand time-series of interest to the supply chain management.

In Figure 8, Figure 9, Figure 10 and Figure 11 we can see graphically the demand for various products. From these figures we can quickly see that, although there are some patterns in the demand, there is also much chaotic behavior, which makes it very difficult for the manufacturer to get accurate forecasts. The first and 10th product have a sight positive trend, the 50th product has a slight negative trend and the 100th product does not seem to have any trend. In all of the 4 series, there is no detectable seasonal pattern and even the large variability of the series changes with time. Variability of the series is increasing with time in the first product and it is decreasing with time in the 100th product and seems stable in the other two.

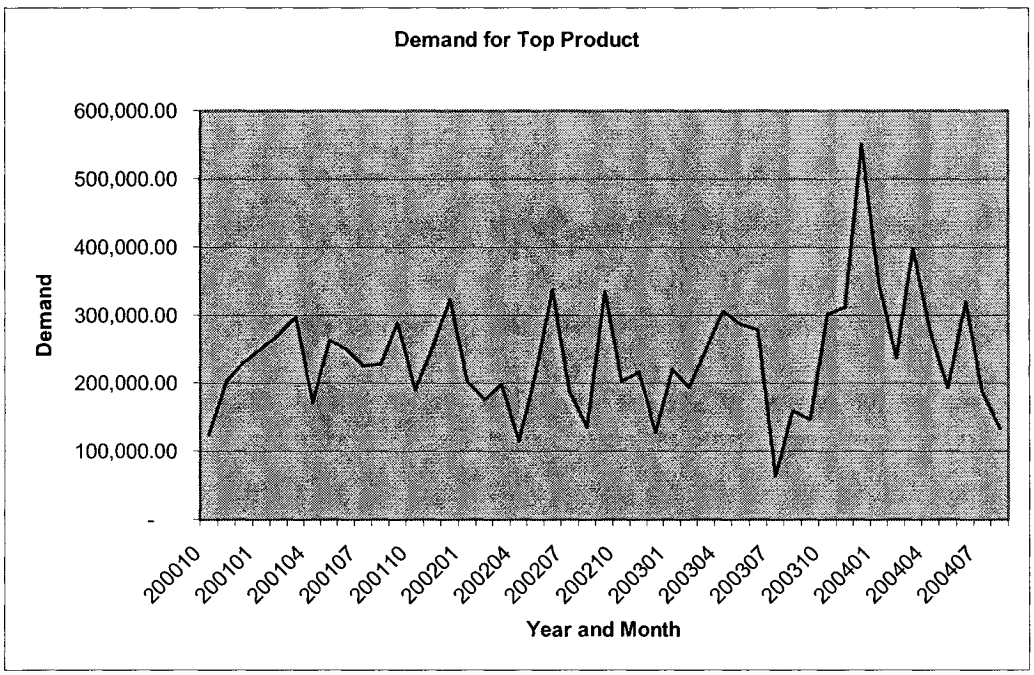

Figure 8 - Demand for the Top Product

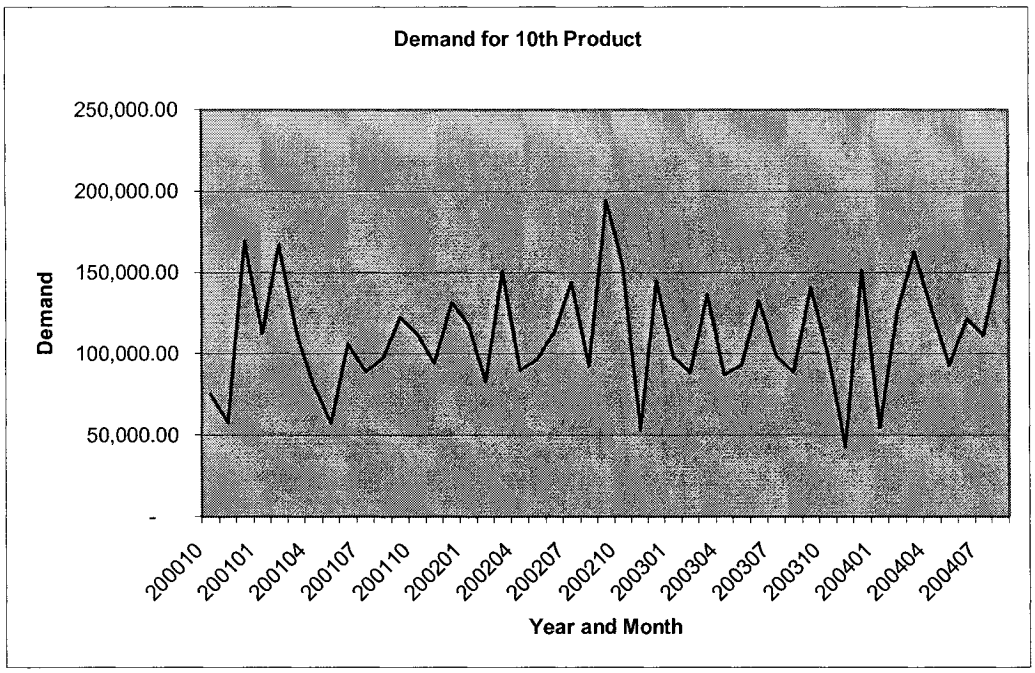

Figure 9 - Demand for the 10th Product

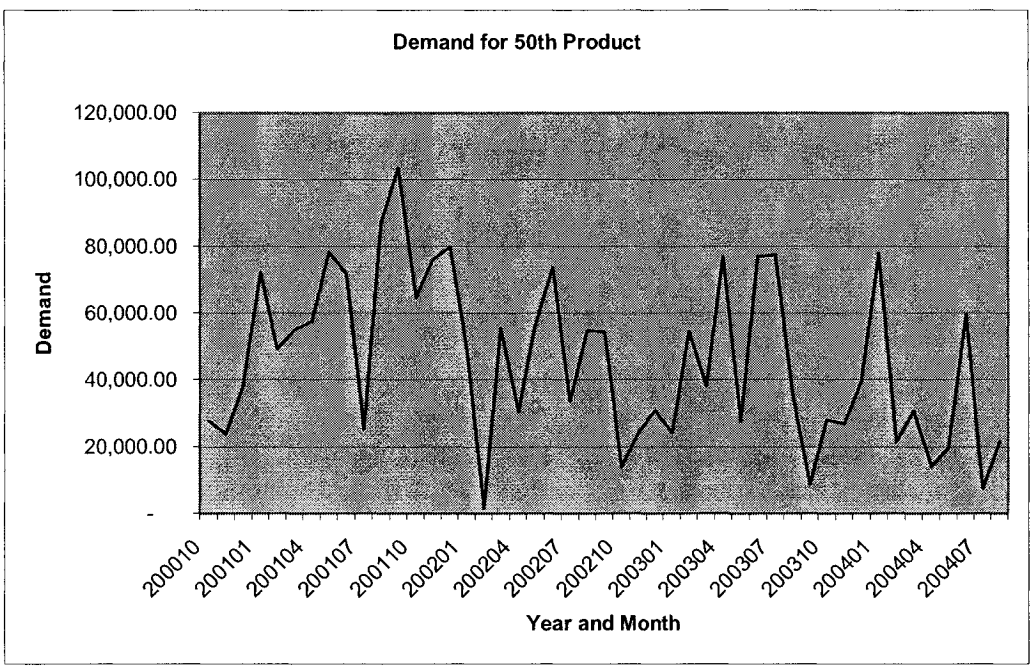

Figure 10 - Demand for the 50th Product

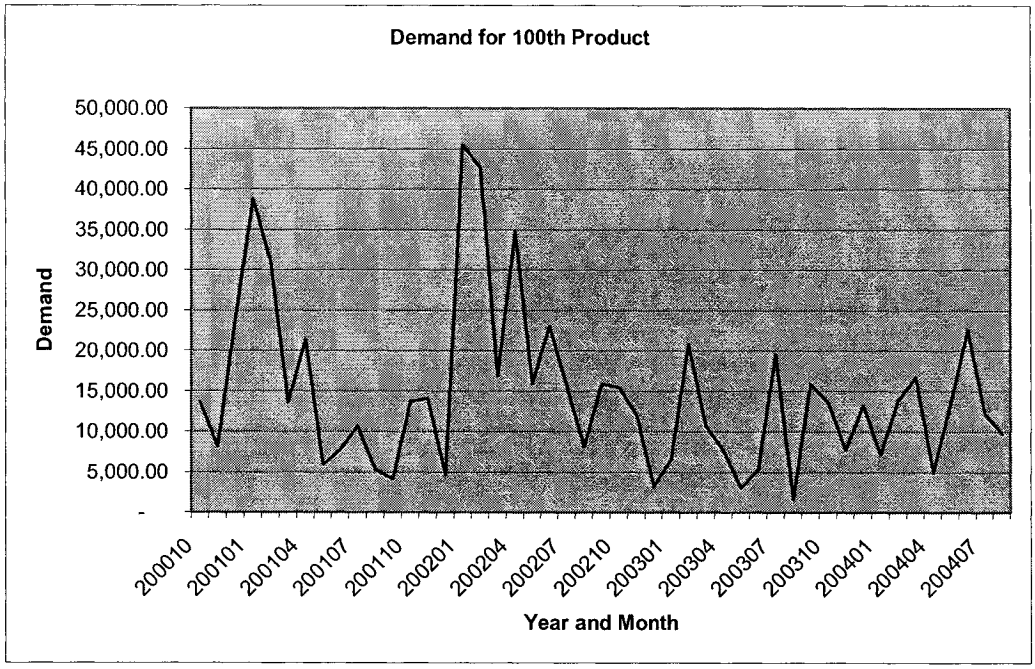

Figure 11 - Demand for the 100th Product

3.1.2 Toner Cartridge Manufacturer

The second manufacturer specializes in generic photocopier and laser printer toner cartridges and other related products. Once again, the geographic scope is of the data is North America. The demand data extracted from the ERP system is from December 1999 to April 2005 and it is aggregated by month that results in 5 years and 5 months of data for a total of 65 periods of data. As with the first manufacturer, only the top 100 products are retained for the experiment, since they are the most important and because this helps keep the experiment feasible. The total number of active products in their system is 3369; so, the 100 products represent approximately 3 percent of all their products. However, these top products account for 38.35% of sales volume for the manufacturer. In Figure 12, Figure 13, Figure 14 and

Figure 15 we can see an overview of the demand of various products. Once again, the time series seem to be quite chaotic with large variation and no clear seasonal patterns. In all of the 4 products, there seems to be somewhat of a downward trend. However, this tread is not always simple, as in the 50th product the pattern seems to be an S shape of demand decreasing, increasing and then decreasing again. The variability of all the 4 series also seems to decrease with time and once again the 50th product shows a more complex S curve in the variation pattern.

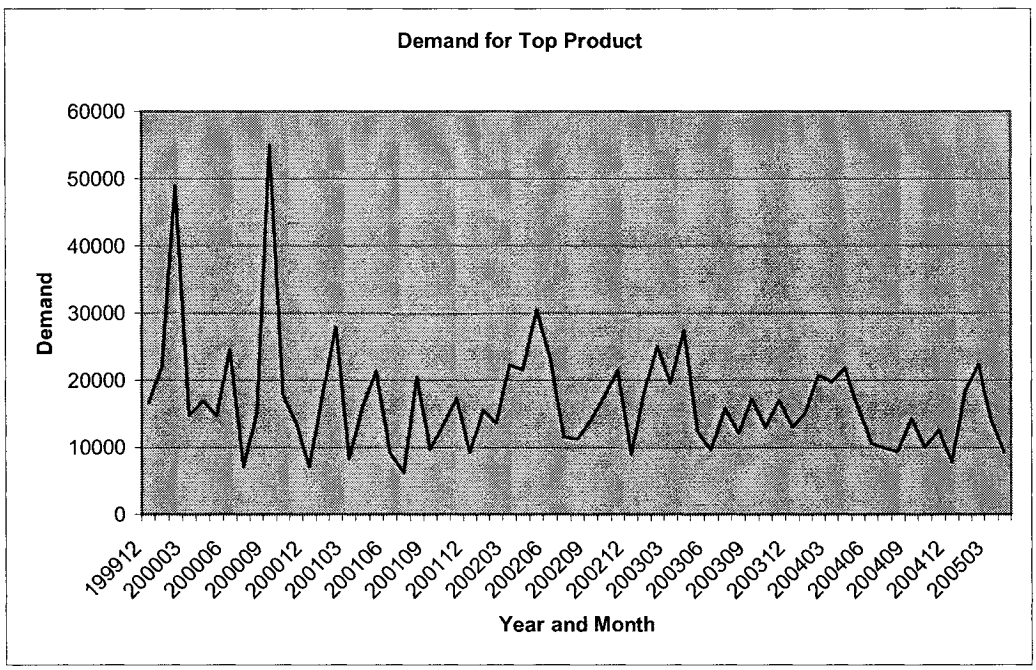

Figure 12 - Demand for Top Product

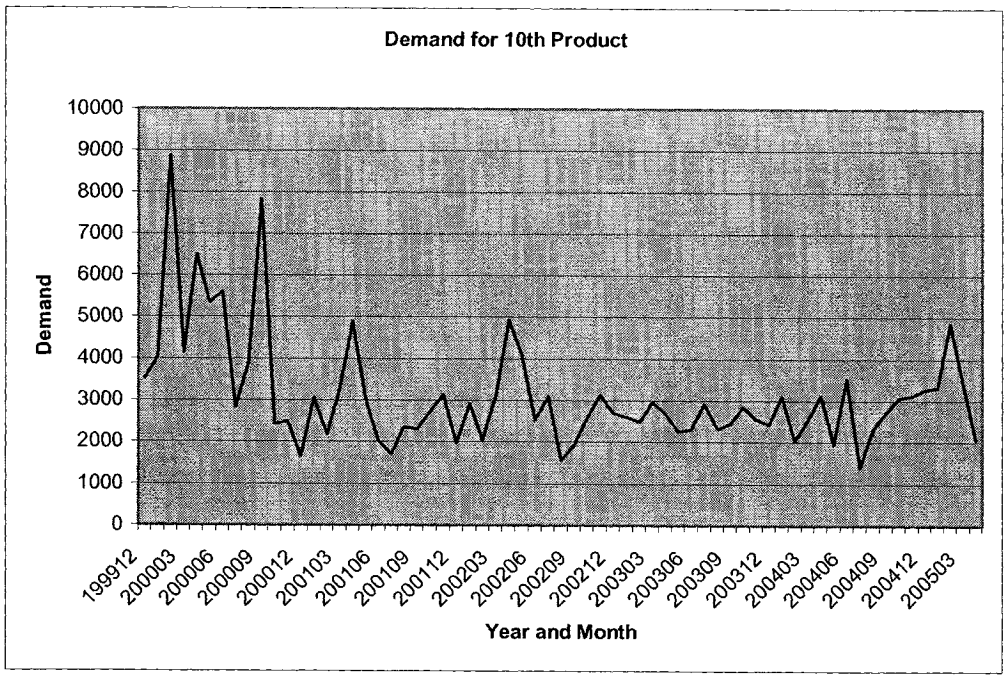

Figure 13 - Demand for 10th Product

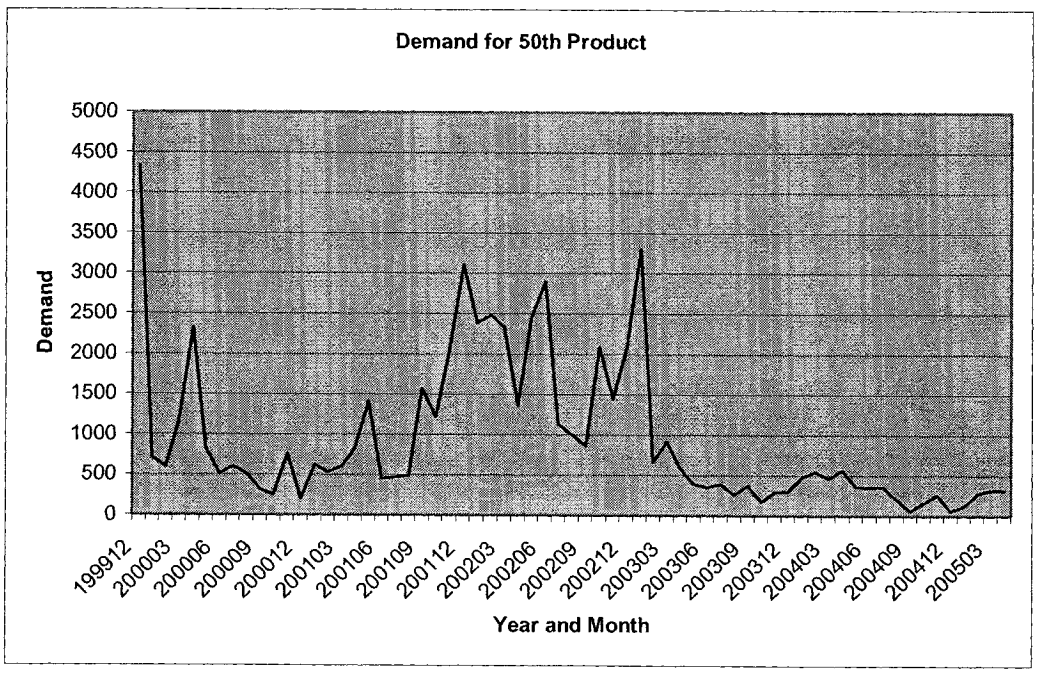

Figure 14 - Demand for 50th Product

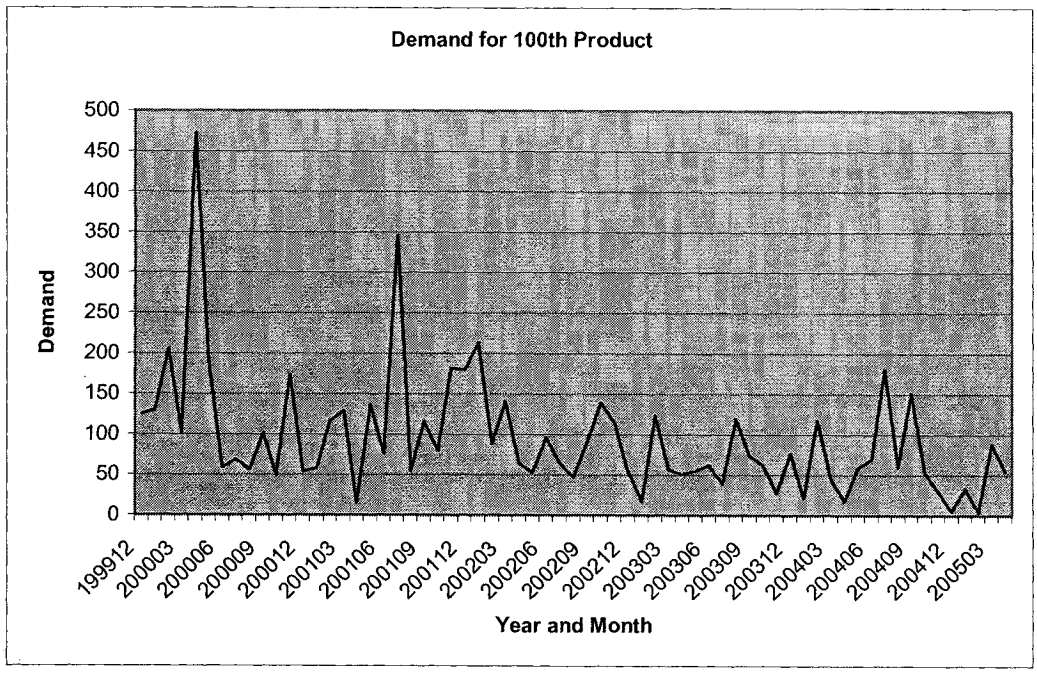

Figure 15 - Demand for 100th Product

3.1.3 Statistics Canada Manufacturing

To add validity to the experiment, we will also include time series data about manufacturers demand as collected by Statistics Canada in the Monthly Survey of Manufacturing (code 2101). More specifically, the data source is the manufacturers' monthly new orders by North American Industry Classification System (NAICS) referred to as table 304-0014 (Statistics Canada 2005). There are several reasons why this dataset is of interest for this experiment.

These are historical data and they cover a large part of the manufacturing sectors in the Canadian economy. Although these data are aggregated for an industry and thus may present different patterns than the individual product data we are using from the two above manufacturer's, it is a good way to experiment on a variety of demand patterns across a large

number of industries which would be otherwise impractical. It will also give us some information about the quality of certain techniques for planning at the aggregate corporate or industry level, which may also be useful for practitioners, even though this is not a primary focus of our study. Additionally, since this dataset is publicly available, it can be used by others to reproduce the results observed in the experiment.

Statistics Canada manufacturing dataset provides "new order" observations starting in January 1992, and at the time of this research, was available up to April 2004. This represents 12 years and 4 months of data or 148 periods. There are 214 industrial categories and to provide a balanced set for the experiment, we selected a random 100 categories. These 100 categories are presented in Appendix I. In Figure 16, Figure 17, Figure 18 and Figure 19 we can see an overview of the demand of various industry categories. The patterns in the Statistics Canada data are different than those of the individual manufacturers because of its aggregate nature and lengthy data collection time span. All four examples show a clear increasing trend and the variability is much lower than that of the individual manufacturers because it is aggregate data. There are some clear sort term cyclical patterns such as in the non-ferrous metal and Urethane series and longer-term cycles in the Sawmill series. Variability is also more stable than in the previous series with a slight overall increase in variability for all the series and slight increase in variability throughout and then a slight decrease at the end of the Sawmill series.

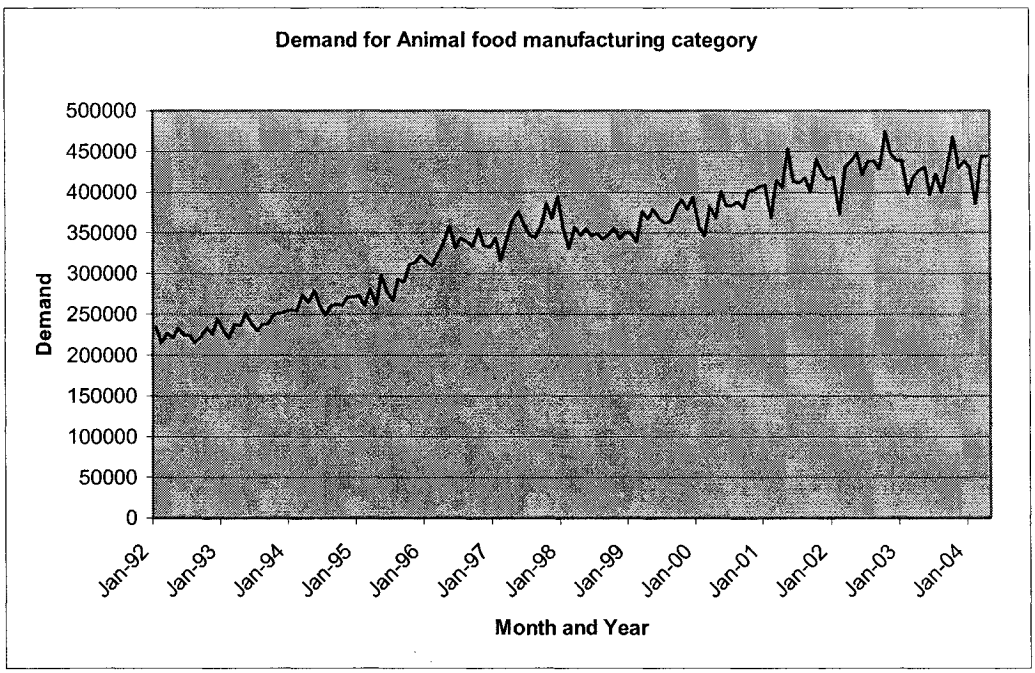

Figure 16 - Demand for Animal food manufacturing category

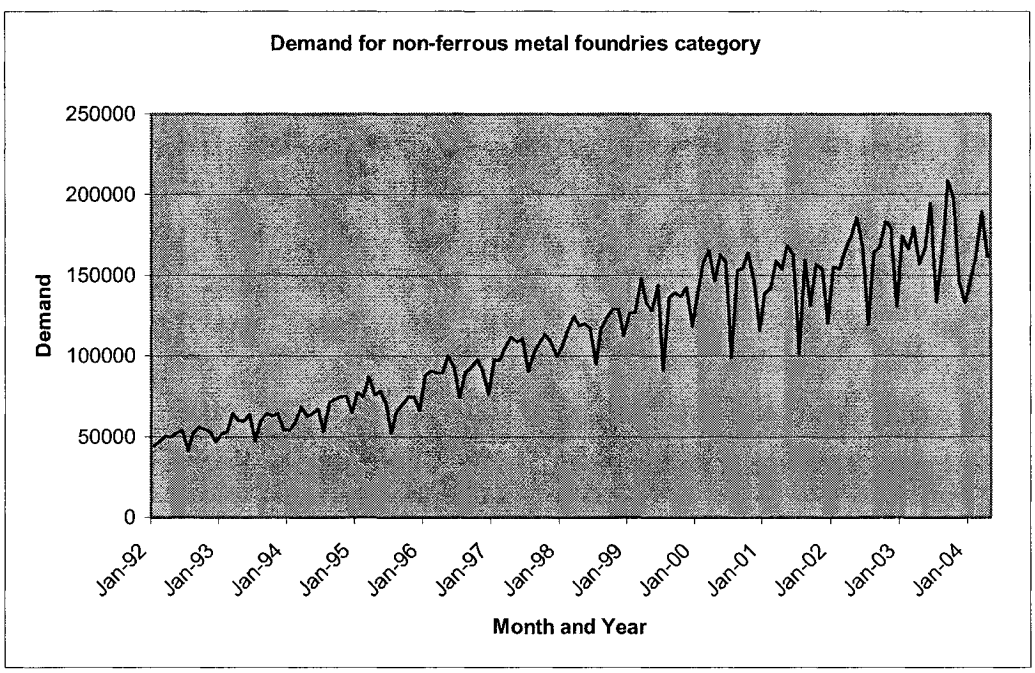

Figure 17 - Demand for non-ferrous metal foundries category

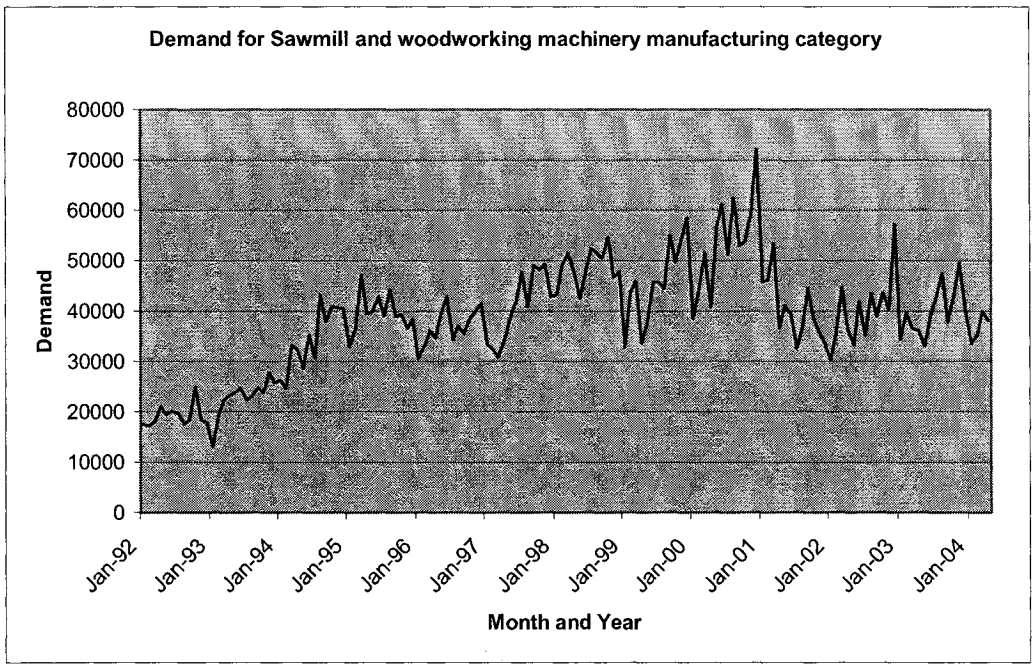

Figure 18 - Demand for Sawmill and woodworking machinery man. cat.

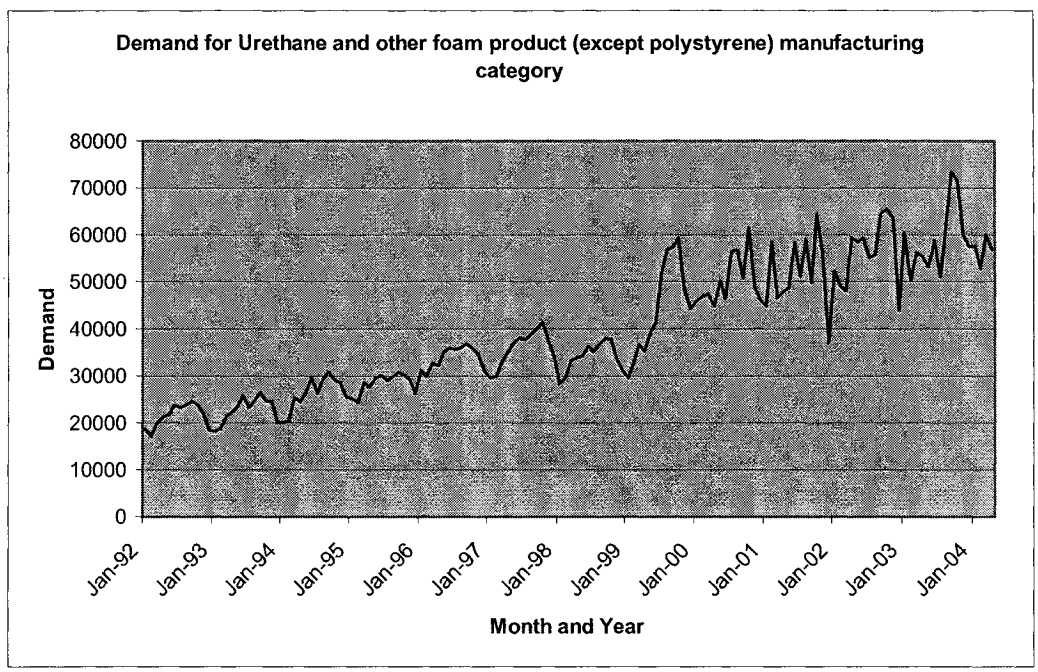

Figure 19 - Demand for Urethane and other ... man. cat.

The 3 selected data sets for the current experiment represent both data at the product level which practitioners must work with in everyday operations as well as data that represents patterns across many parts of the Canadian economy. The patterns in these data sets are diverse and very noisy and distorted and this is in line with both theoretical and empirical observations of demand patterns at the manufacturer's end of the supply chain. The series seem to be noisy because there are a important portion of the variations in the demand that do not seem to follow any pattern that can be understood or predicted. The series seem to be distorted because they do not resemble any demand patterns that would be expected at the customer end of the supply chain, such as seasonal cycle. These three datasets that represent a total of 300 timeseries will be the empirical observation sources for all of our subsequent experiments.

4 Research methodology

This research project will use an experimental research design and methodology. As previously mentioned, we have selected a representative set of traditional forecasting techniques used by practitioners and identified by researchers as the best overall forecasting techniques. Forecasts made by this group are considered the control group Table 3. The other set of techniques are artificial-intelligence based and the forecasts they produce represent the treatment group Table 3.

Table 3 - Overview of Control and Treatment Group

CONTROL GROUP Traditional Techniques	TREATMENT GROUP Artificial Intelligence Techniques
• Moving Average • Trend • Exponential Smoothing • Theta Model (Assimakopoulos & Nikolopoulos 1999) • Auto-Regressive and Moving Average (ARMA) (Box and al. 1994) • Multiple Linear Regression (Auto-Regressive)	• Neural Networks • Recurrent Neural Networks • Support Vector Machines

To compare the two groups, every technique from each group will be used to forecast the demand one month into the future for all of the 100 series for the 3 datasets previously identified. This will result in a series of a maximum 4700 forecast points for the chocolate

manufacturer, 6500 for the toner cartridge manufacturer and 14800 for the Statistics Canada dataset for every technique tested, however, since all forecasting techniques require past data to make a forecast into the future, there will be a predetermined startup period which will slightly reduce the number of forecast observations.

Additionally, as will be detailed later, the demand time series will be formally separated into a training set and testing set. This is very important especially for the artificial intelligence techniques. Since they are universal approximators and can learn any function to an arbitrary precision, they can easily memorize the demand series and thus have extremely good forecasting accuracy. Unfortunately, because these series have both patterns and noise, such high level of precision in learning will result in learning the real patterns as well as the noise. This effectively means that a technique will have memorized the demand and although it will have very good forecasting accuracy on the set it learnt from, it will have very poor forecasting when used in the future since it has learnt some noise which by definition will not occur again in the same pattern.

Therefore, we must have both a training and a testing set, the training set is used for teaching the techniques about the demand patterns and the testing set is used to estimate how well the forecast will generalize in the future. Performance on the training set is of interest to understand what the various techniques are learning. However, all performance related measurements that are used for testing the hypothesis are done on the testing set since the objective is to identify the best technique that will generalize into the future.

The main performance measure which will be used to test the proposed hypothesis will be the Absolute Error (AE) measure. For every forecast data point for which we have actual data, we can calculate the error and convert it to an absolute error. This will result in a series of absolute error values for a specified forecasting technique. However, to make the absolute error comparable across products, we must normalize this measure. Thus the absolute error will be expressed in terms of standard deviations of the training set. This series of Normalized Absolute Error (NAE) can then be compared with another series using a t-test to determine if there is a statistical difference in the error of the techniques, which tells us if there is a difference in forecasting performance. The absolute error can also be summarized for a specific product demand as the Mean Absolute Error (MAE) for a high level overview of the forecasting performances.

5 Experiment Implementation

To test the proposed hypothesis, we will have to execute all of the forecasting algorithms on the demand series from the 3 datasets. The first step to the implementation of this experiment is the preparation of the data and the separation into training and testing sets. The second step is the implementation of all the forecasting techniques selected so that the forecasts may be executed. Thirdly, the forecasting of the products and the collection of statistics must be automated, because of the large amount of work required to manually do this is infeasible. All of the data processing and forecasting will be performed in the MATLAB 7.0 environment (MathWorks 2005c) because of the many mathematical and statistical algorithms available as well as the flexible research environment it offers.

5.1 Data preparation

The data are prepared in an intermediate file format, which will be used to load them into the MATLAB environment for processing. For example, the product demand of the chocolate manufacturer for the 100 top products over 47 months is prepared in a matrix in which each product is represented by a row going from top to bottom and each month is represented by a column from left to right. A small portion of the table is presented in Table 4. The first row of this table represents the historical demand for product one, the first demand is of 124,397.61 Kilograms for October 2000, the next is of 203,095.82 Kilograms for November 2000. This matrix can then be loaded into MATLAB where all the data processing will occur.

Table 4 - Sample segment of Product Demand Matrix

Product	2000/10	2000/11	2000/12	2001/01	2001/02
Product 1	124,397.61	203,095.82	230,538.13	251,221.93	270,454.23
Product 2	413,090.78	223,235.32	444,497.48	206,345.84	234,706.64
Product 3	180,076.03	201,730.51	316,389.49	100,506.92	204,388.56
Product 4	153,767.68	144,174.22	230,243.30	118,296.80	136,281.72
Product 5	104,308.02	123,748.97	246,690.55	166,867.43	187,651.01

For accurate performance testing, we must separate the data into training and testing sets. The main objective of dividing the data into these two sets is to have enough data in the training set for the techniques and to have enough data in the test set to appropriately verify the quality of the model. Since most data sources are limited, a much larger portion of the data commonly is used for training than for testing. Our time series are very small, so we will use 80% of the time series data to train and 20% of the data to test. The 80% training set is important since it allocates a large portion of the data to training considering the small size of the time series and the 20% test set is an adequate size for verifying the model quality.

To illustrate, in the chocolate factory dataset, the training set contains 80% of the data, thus 38 months of demand and the testing set will contain the other 20% of the data that is 9 months of demand. This represents data from October 2000 to November 2003 for the training set and December 2003 to August 2004 for the testing set. In terms of total data points from the testing set that will be compared in the t-test, this represents 900 forecast observations to compare between each forecasting technique.

When a forecasting algorithm requires windowed data, such as Multiple Linear Regression, Neural Networks and Support Vector Machines, the procedure is as follows. A window defines how much past data is used to predict a future demand. For example a window size of 3 months could be defined thus indicating that this current month's data and the data from the previous 2 months is used to predict next month's data. This is illustrated in Figure 20. Taking observation 3 (o3) as an example, the demand at time period 6 (t6) is modeled as a function of the demand at time period 3 (t3), 4 (t4) and 5 (t5). This example simulates being at time period 5 (t5) and using the current demand (t5) and past 2 periods of demand (t3, t4) to predict future demand (t6). However, since this is a simulation, the learning algorithms can be taught the correction function by presenting the known demand of the simulated future period (t6).

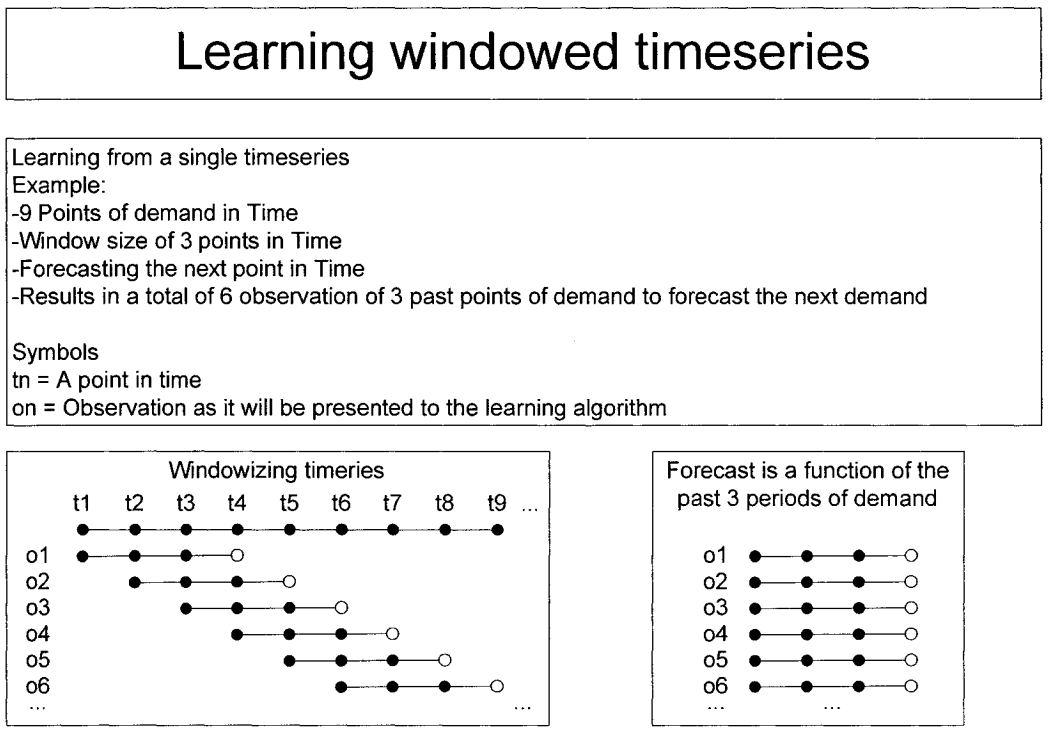

Figure 20 - Learning from a windowed timeseries

5.2 Forecasting Implementations

A broad overview of forecasting techniques that will be used in this experiment were previously presented and we will now present more specific details on their implementation. For some of the simpler forecasting techniques such as the Moving Average, Trend, and Exponential Smoothing, which all have one parameter that needs to be set, we can identify this parameter by searching for the parameter that provides the best forecast on the training set. This will be referred to as the automatic version of the technique. However we also include a comparison implementation of the forecasting method that uses a constant parameter term that seems reasonable to help identify if the automatic version is performing well and identify any anomalies with the automatic procedure. However, even an extremely simple exponential smoothing algorithm presents some implementation problems.

Although we have seen that previous research has identified Exponential Smoothing as a very good approach, there are several potential implementations of it which make implementation and comparison with previous results difficult. For example, the exponential smoothing approach can use the initial value of the series as a starting value or the average of the series as a starting value. The exponential smoothing implementation in MATLAB Financial Toolbox (MathWorks 2005a) and in Excel use the first value as the initial value, but, other implementations such as SPSS use the series average as the starting value. When combining this with automatic parameter selection, a large difference in results occurs as will be seen later. For this reason we implement both versions of exponential smoothing and the same thing will apply in the implementation of the Theta model. In some algorithms, such as the Theta model and the ARAMAX model, the automatic parameter selection procedure is already defined as

part of the algorithm. The main purpose of the Multiple Linear Regression model is to provide a linear benchmark for all of the auto-regressive type models such as the Neural Networks and Support Vector Machines.

Below we present the final list of forecasting algorithms that will be used in the experiments

- Moving Average 6 Months
- Moving Average (Automatic)
- Trend 6 Months
- Trend (Automatic)
- Theta Initial
- Theta Average
- Exponential Smoothing 20% Initial
- Exponential Smoothing (Automatic) Initial
- Exponential Smoothing 20% Average
- Exponential Smoothing (Automatic) Average
- ARMA
- MLR
- MLR SuperWide
- ANN LMBR
- ANN BPCV
- ANN LMBR SuperWide
- ANN BPCV SuperWide
- RNN LMBR
- RNN BPCV
- RNN LMBR SuperWide
- RNN BPCV SuperWide

- SVM CV
- SVM CV SuperWide
- SVM CVWindow
- SVM CVWindow SuperWide

Legend:

- ARMA – Auto-Regressive Moving Average
- MLR – Multiple Linear Regression
- ANN – Artificial Neural Networks
- RNN – Recurrent Neural Networks
- BPCV – Back Propagation with Cross Validation based early stopping
- LMBR – Lavenberg-Marquardt training with Bayesian Regularization
- SVM – Support Vector Machines
- CV – Cross Validation
- CVWindow – Cross Validation based on a sliding Window

5.2.1 ARMA

The ARMA model combines both Auto-Regressive forecast and a Moving Average forecast. The selection of the lag used in the auto-regression portion and the lag used in the moving average portion is optimized to minimize the error. This functionality is provided by the MATLAB GARCH Toolbox (MathWorks 2005b) which implements Generalized Autoregressive Conditional Heteroscedasticity models. Within the toolbox, the ARMAX model is provided. This model which is a super-set of the ARMA model since it is a generalization that supports additional series data to be added via regression into the model, hence the X in the name. The ARMAX model is optimized to minimize the error using the

MATLAB Optimization Toolbox (MathWorks 2005e). Only the ARMA part of the ARMAX model is used in the current experiments.

5.2.2 Theta Model

To ensure proper reproduction of the Theta model (Assimakopoulos and Nikolpoulos 2000), one of the researchers, Konstantinos Nikolopoulos has reviewed the details and the forecasting results of the implementation. The Theta model is a flexible forecasting strategy. However, our implementation follows what was used in the M3 forecasting competition. The data series is decomposed by extracting the linear trend and the remaining patterns are doubled. On this subsequent decomposed and doubled series, exponential smoothing is optimized to minimize the error on the training set. The two individual series, the linear trend and the optimized exponential smoothing on the decomposed series are recombined by an average of the two. However, as will be seen in the results, depending on the selection of the exponential smoothing initialization, the results vary significantly, and the researchers do not provide specific guidance on this topic only mentioning that both methods are valid. Therefore, to have a more complete experiment, we will implement both versions of the Theta, the initial value as the initialization and the average of the training set as the initialization.

5.2.3 Neural Networks details

Using a window size of 5% of the training set data for the regular timeseries data models, we already have a ratio of 1 inputs to 20 observations. Thus once the hidden neurons are added, the neural network is very powerful compared to the number of observations that will be used

to train it. Therefore, to provide non-linearity and additional modeling power, but not too much, we create one hidden layer which contains 2 neurons with non-linear transfer functions. Even though there are only 2 hidden layer neurons and considering the inputs space and small datasets, if the neural networks uses all of these weights, it will be definitely overfitting the data. The total number of weights for a neural network with one hidden layer can be calculated as follows.

Using the following variables:

p = number of periods in the data
w = window ratio
h = hidden layer neurons
b = bias (always 1)
o = number of outputs

We can calculate the total number of weights as:

$$\text{Total Weights} = p \cdot w \cdot h + b \cdot h + h \cdot o + b \cdot o$$

Therefore for the current implementation, the number of weights will always be:

$$\text{Total Weights} = p \cdot w \cdot 2 + 1 \cdot 2 + 2 \cdot 1 + 1 \cdot 1$$

$$\text{Total Weights} = p \cdot w \cdot 2 + 5$$

And the Observations to Weights ratio is:

$$\text{Observations to Weights} = \frac{p \cdot (1 - w)}{(p \cdot w \cdot h + b \cdot h + h \cdot o + b \cdot o)}$$

Therefore, for the chocolate manufacturer dataset, Observations to Weights ratio is:

$$\text{Observations to Weights} = \frac{38 \cdot (1 - 0.05)}{(38 \cdot 0.05 \cdot 2 + 1 \cdot 2 + 2 \cdot 1 + 1 \cdot 1)}$$

$$\text{Observations to Weights} = 4.10$$

As with linear regression where at least 10 observations per variable is desirable, there should be a minimum of 10 observations for each weight. This estimation varies based on the

expected complexity of the pattern being modeled. For example, the more complex the expected pattern or the more noise, the more observation to weights required. Thus an observations to weights ratio of 4.1 is very low. However, because timeseries forecasting is a function of the past information, we felt that the window size must include 2 or more periods. For the current research project, since our smallest data set is 47 periods, thus 38 for the training set, a window size of 5% represents 2 past periods. If we reduce the window size anymore, this model will only have 1 period in the window and we would not have a very meaningful model. For our largest dataset, there are 148 periods, 118 for the training set, thus representing a window of 6 periods.

Each neuron sums all of its inputs and then processes them through a transfer function, which can permit non-linear modeling and which can squash the sum to a normalized range. The transfer function we used in the hidden layer is the tan-sigmoid function which does non-linear scaling from an infinite range to values between -1 and 1 and the output layer transfer function is linear, meaning that the outputs are not modified, which consequently permits outputs outside the range of -1 to 1. Additionally, each neuron contains a bias input which is a constant of 1. The inputs for each neuron are the input signals coming from either the input variables or the results of a previous neural network layer, multiplied by the weight of the connection.

The relevant features of the supply chain demand modeling neural network are displayed in Figure 21. In this figure, the sum is represented by the Greek capital letter sigma (Σ), the tan-sigmoid transfer function is represented by a horizontal axis with an S shape crossing it ($\mathcal{F}$) and the linear transfer function is represented by a horizontal axis with a diagonal line crossing

it ($\nearrow$). All of the inputs to the neural network, as well as the output are individually scaled between -1 and 1 to ensure that they are within the appropriate range for neural network training. The final results are then unscaled to permit comprehensible analysis and usage.

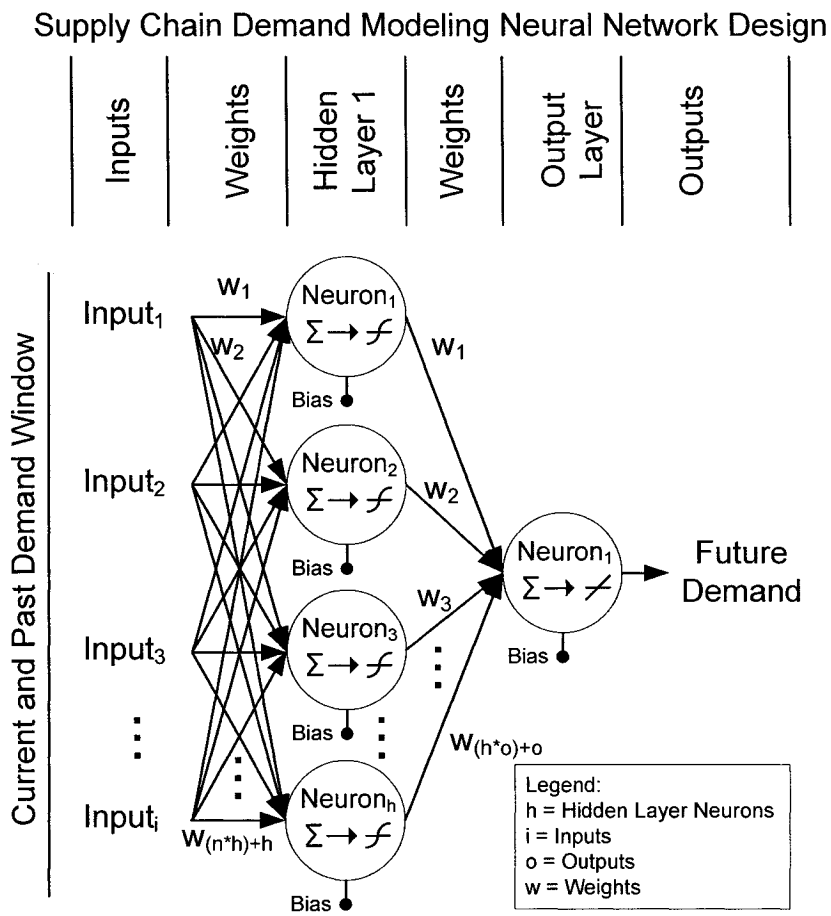

Figure 21 - Supply Chain Demand Modeling Neural Network Design

The following is a descriptive example for Figure 21. A selected window size of 3 months would result in 3 inputs to the network. Setting the number of hidden layer neurons at 2,

would result in a total of 6 weights (3*2) and 2 bias weights in the first half of the network. In the second half of the network, there would be 2 weights connecting the hidden layer and the output layer and there would be a bias on the final output neuron. Additionally, the hidden layer neurons have a non-linear transfer function and the output later neuron has a linear transfer function.

The first implementation of Neural Networks is based on the traditional backpropagation algorithm. The structure of neural networks must be defined in advance by specifying such parameters as the number of hidden layers and the neurons within each hidden layer. Other settings that must be defined relate to the learning algorithm, these include the learning rate and the momentum of the learning rate.

One parameter of interest is the learning rate since setting a constant learning rate for the training session is not desirable because the ideal learning rate may change based on the current progress of the networks learning. An adaptive variable learning rate training algorithm has been developed which adjusts the learning rate for the current learning error space (Hagan and al. 1996). This algorithm tries to maximize the learning rate subject to stable learning, thus adapting to the complexity of the local learning error space. For example, if the decent path to the lowest error is straight and simple, the learning rate will be high. If the decent path is variable, complicated and unclear, the learning rate will be very small to permit more stable learning and avoid jumping around in the learning space. In addition to the variable learning rate, our first neural network learning algorithm also includes the momentum (Hagan and al. 1996).

To help the neural network stop training before it overfits the training set to the detriment of generalization, we use a cross validation set for early stopping. This cross validation set is an attempt to estimate the neural network's generalization performance. As previously presented, based on the amount of data available, we have defined the cross validation set as the last 20% of the training set. This set is removed from the training set and is verified after every training epoch. An epoch is a single cycle over all of the training data. The error on the cross validation set will decrease as the network starts to learn general patterns and then will increase as the network starts to memorize the training set. Thus, the weights that resulted in the lowest error rate on the cross validation set are identified as the neural network model that provides the best generalization performance.

An example graph of the training and cross validation set errors is presented in Figure 22 where we see the training set error as a dark shaded line and the cross validation set error as a light shaded line. The y-axis represents the error and the x-axis represents the epochs, so the Figure 22 presents a visual representation of the error minimization as the neural network learns through the epochs. The example presented is currently at about epoch 145, and we can see that the cross validation set error was at its lowest point at around epoch 65. Therefore, because the cross validation set error increases after that, this suggests that the neural network is presumably overfitting.

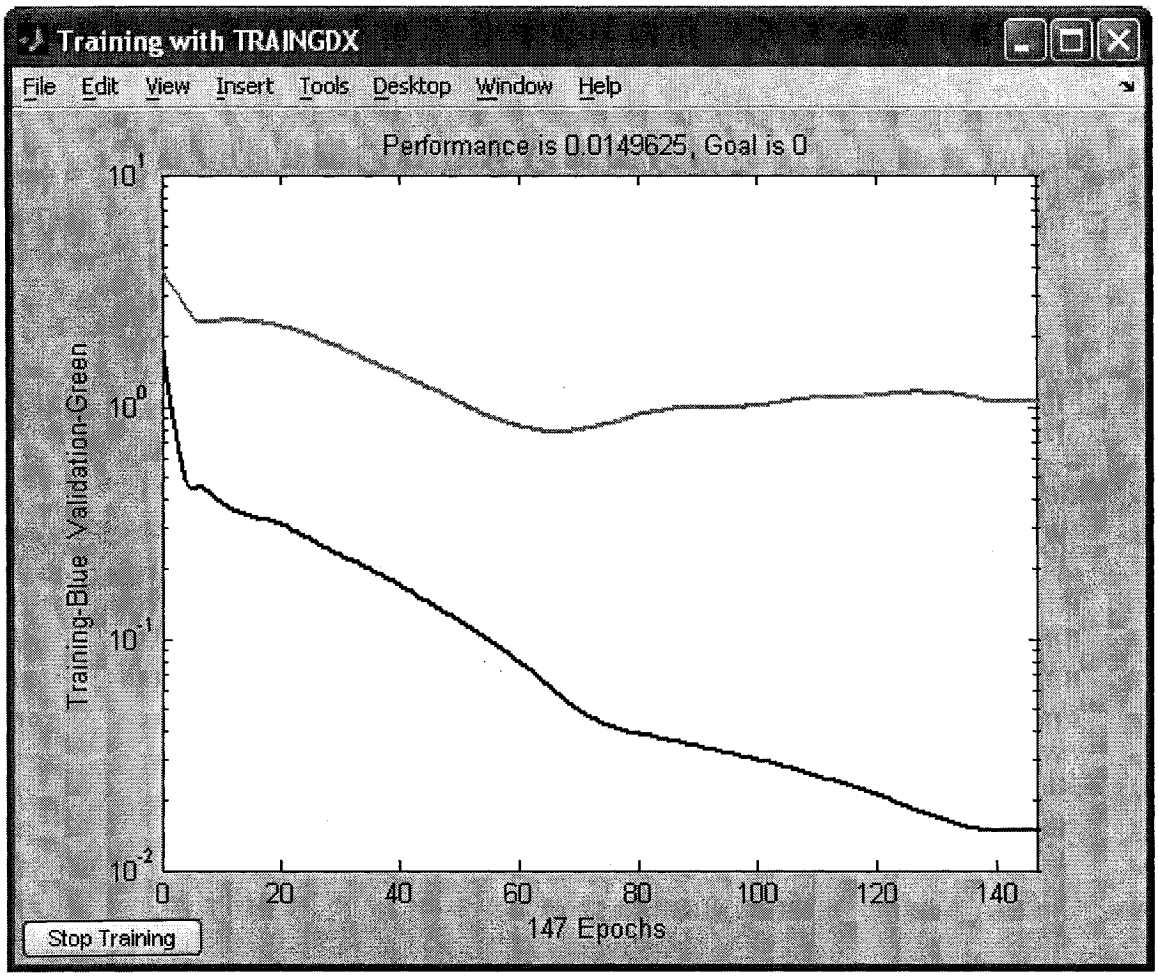

Figure 22 - Example Neural Network Training and Cross Validation Errors

In addition to testing the backpropagation learning algorithm with cross validation early stopping, we will use a faster training algorithms as well as a framework that will increase the generalization of the model. In particular, we will use the Levenberg-Marquardt algorithm (Marquardt 1963) as applied to Neural Networks (Hagan and Menhaj 1994; Hagan, Demuth et al. 1996) for adjusting the weights so that learning of patterns may occur. This algorithm is

one of the fastest training algorithms available with training being 10-100 times faster than simple gradient decent backpropogation of error (Hagan and Menhaj 1994).

The Levenberg-Marquardt neural network training algorithm is further combined into a framework that permits estimation of the network's generalization by the use of a regularization parameter. Neural network performance measures typically measure the error of the outputs of the network, such as the means squared error (MSE). However, a regularization performance function which includes the sum of the weights and biases can be used instead, combined with a regularization parameter which determines how much weight is given to the sum of weights and bias in the formula (msereg = γ mse + (1 - γ) msw). This regularization parameter permits the control of the ratio of impact between reducing the error of the network and the number of weights or power of the network which means that one can be less concerned with the size of the neural network and control the effective power of it directly by the use of this parameter.

The tuning of this regularization parameter is automated within the Bayesian framework (MacKay 1992) and, when combined with the Levenberg-Marquardt training algorithm, results in high performance training combined with a preservation of generalization by avoiding overfitting of the training data (Foresee and Hagan 1997). Not only does this algorithm help eliminate overfitting of the target function, it also provides an estimate of how many weights and biases are being effectively used by the network. Larger networks should result in approximately the same performance since regularization results in a trade off between error and network parameters which is relatively independent of network size.

All Neural Network modeling and training is performed in MATLAB 7.0 and MATLAB's Neural Network Toolbox (MathWorks 2005d). An example of a Levenberg-Marquardt with Automated Bayesian Regularization training session is presented in Figure 23 where we can see that the algorithm is attempting to converge the network to a point of best generalization based on the current training set. Even though this particular network has 256 weights, the algorithm is controlling the power of the neural network at effective number of parameters of about 44. The network could further reduce the error on the training set (Sum of Squared Error: SSE) since it could use all 256 weights. However, it has determined that using more than the 44 weights will cause overfitting of the data and thus reduced generalization performance.

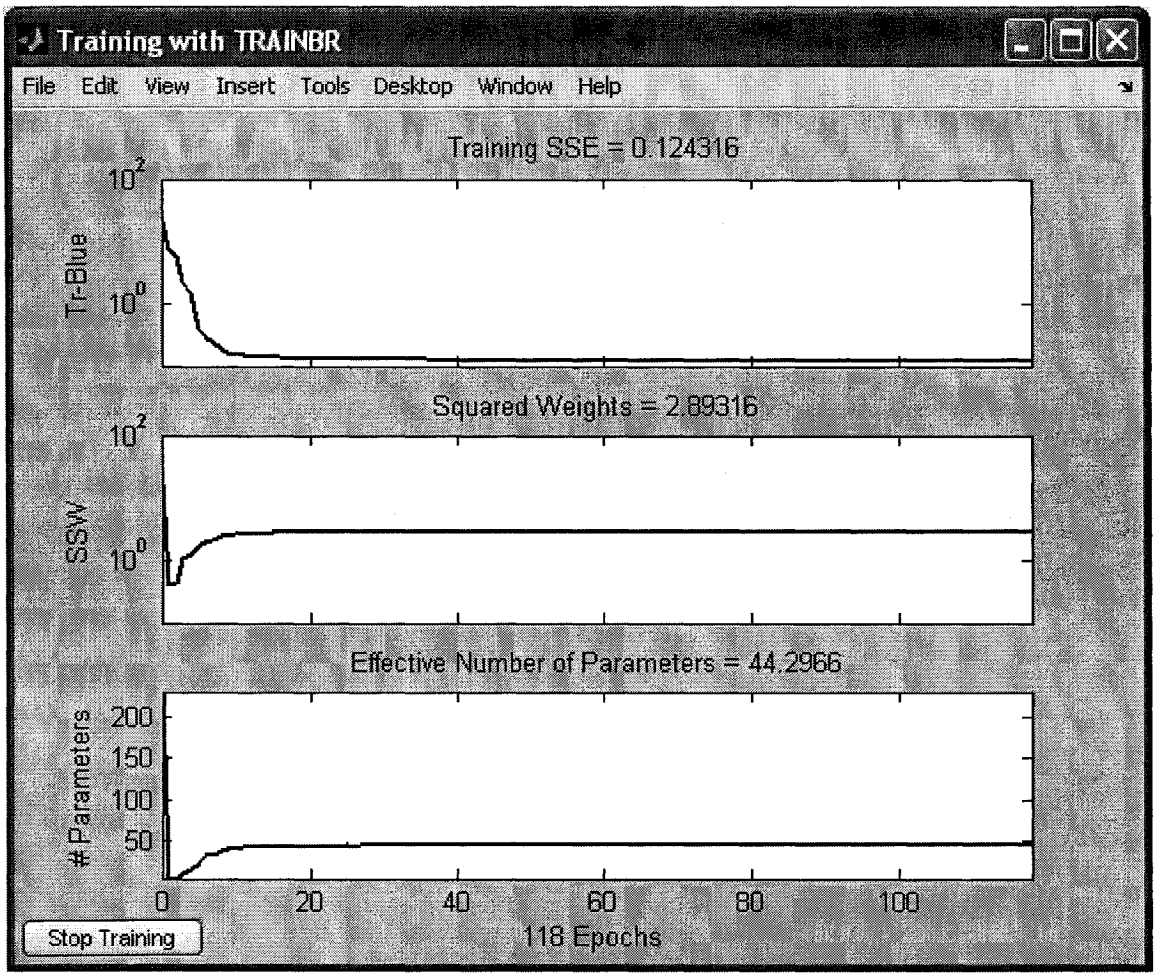

Figure 23 - Example Levenberg-Marquardt Neural Network training details.

Compared to the early stopping based on a cross validation set, the Levenberg-Marquardt with Automated Bayesian Regularization training algorithm is superior especially for small datasets since separating out a cross validation set it not required because the algorithms looks at all the training data. However, the disadvantage is that it does not sample the data in a time-series sequential method, therefore potentially identifying future patterns that generalize well into the past, but that do not generalize well into the future. Generalizing quality for the past is not

useful since we already know the past; it is only generalization into the future that we are interested in.

5.2.4 Recurrent Neural Networks details

The recurrent neural network architecture is the same as the above described feedforward architecture, except for one distinguishing difference. There are recurrent network connections within the hidden layer as presented in the subset architecture in Figure 24. This architecture is known as an Elman Network (Elman 1990). The recurrent connections feed information from the past execution cycle back into the network. This permits a neural network to learn patterns through time. Thus the same recurrent networks with the same weights and given the same inputs may result in different outputs depending on the feedback signals currently held in the network. In our experiments we will use the two previously described training methods, the variable learning rate with momentum and early stopping based on cross validation set error and the Levenberg-Marquardt with Automated Bayesian Regularization training algorithm. The addition of recurrent connections also increases the size of the network by the number of hidden layer neurons squared.

Using the following variables:

p = number of periods in the data
w = window ratio
h = hidden layer neurons
b = Bias (always 1)
o = Number of outputs

We can calculate the total number of weights as:

$$\text{Total Weights} = p \cdot w \cdot h + b \cdot h + h \cdot h + h \cdot o + b \cdot o$$

Therefore for the current implementation, the number of weights will always be:

$$\text{Total Weights} = p \cdot w \cdot 2 + 1 \cdot 2 + 2 \cdot 2 + 2 \cdot 1 + 1 \cdot 1$$

$$\text{Total Weights} = p \cdot w \cdot 2 + 9$$

Therefore, for the chocolate manufacturer dataset, Observations to Weights ratio is:

$$\text{Observations to Weights} = \frac{38 \cdot (1 - 0.05)}{(38 \cdot 0.05 \cdot 2 + 1 \cdot 2 + 2 \cdot 2 + 2 \cdot 1 + 1 \cdot 1)}$$

$$\text{Observations to Weights} = 2.82$$

So the Observations to Weights Ratio is even lower than the 4.1 previously identified for the neural network without the recurrent connections which was already low. However, the window size ratio and the number of hidden layer neurons should not be further reduced since they are already at their lowest meaningful levels.

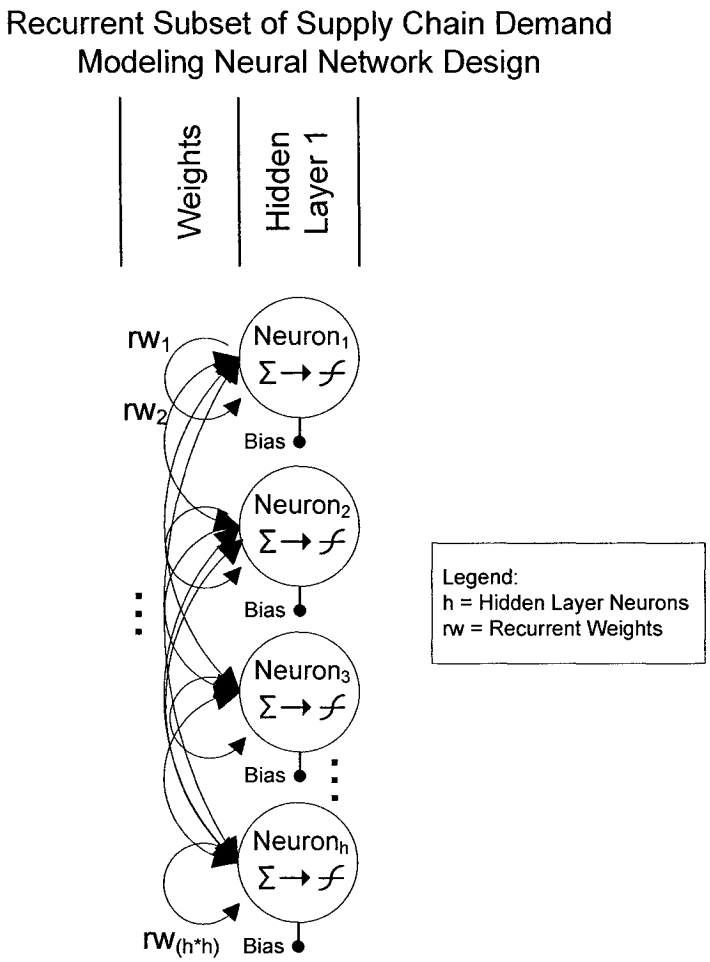

Figure 24 - Recurrent Subset of Neural Network Design

5.2.5 Support Vector Machine details

The support vector machine software implementation selected for the current experiment is mySVM (Rüping 2005) which is based on the SVMLight optimization algorithm (Joachims 1999). The inner product kernel is used and the complexity constant is automatically determined using cross validation procedure.

Two cross validation procedures are tested. The first is a simple 10 fold cross validation that ignores the time direction of the data. Thus, for 10 iterations, $9/10^{th}$ of the data is used to build a model and the remaining $1/10^{th}$ is used to test its accuracy. The second simulates time ordered predictions, called windowed cross validation. This cross validation procedure splits the training data set into 10 parts and the algorithm trains the model using 5 parts and tests on a 6^{th} part. This 5-part window is moved along the data which results in the procedure being repeated 5 times. For example blocks 1-5 are used to train and the model is tested on block 6, then blocks 2-6 are used to train the model and it is tested on block 7, this continues until blocks 4-9 are used to train and the model is testing on block 10.

The error of these five models is averaged and the complexity constant with the smallest cross validation error is selected as the level of complexity that provides the best generalization. The range of a very small complexity constant which does model the data well to a very large complexity constant which overfits the data results in a error curve which permits the minimization of the generalization error. An example error curve for the complexity constant search on a 10-fold cross validation set with a 5 fold sliding window for the complexity constant range between 0.00000001 and 100 with a multiplicative step of 1.1, is presented in Figure 25.

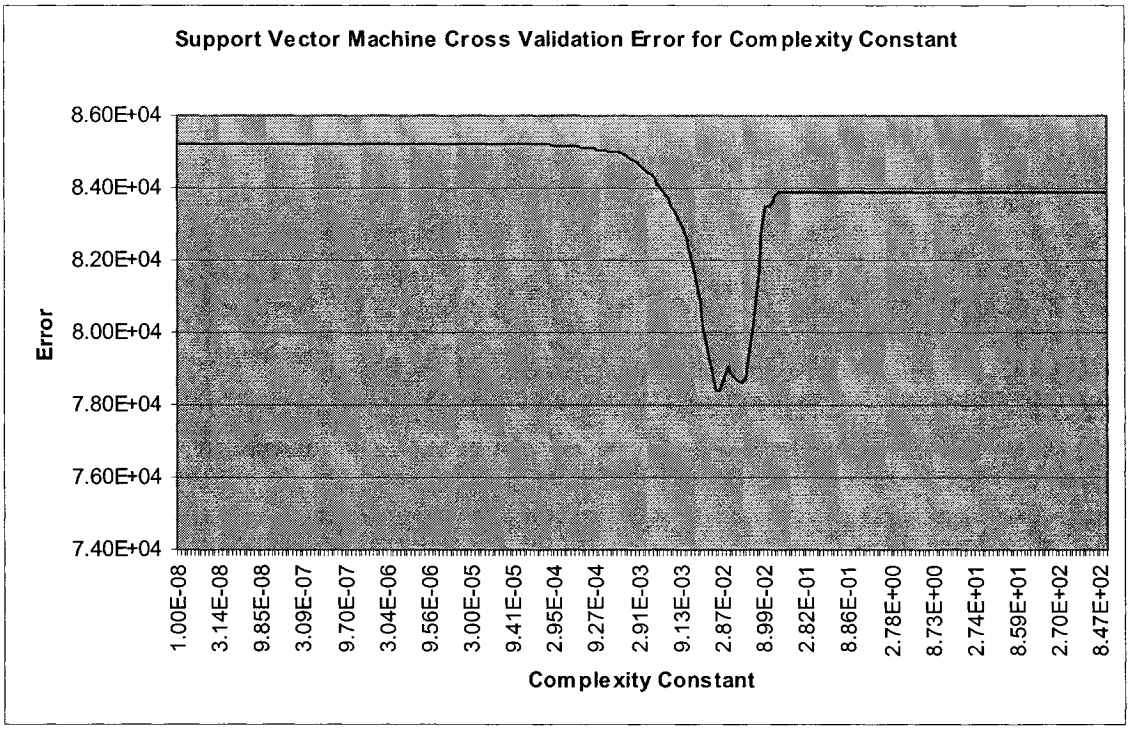

Figure 25 - SVM Cross Validation Error for Complexity Constant

We also present an example of the data underfitting the model with complexity constant 0.00000001, overfitting with 1000 and the optimal estimated generalization fit is with a complexity constant of 0.012154154 in Figure 26. In this diagram, we can see that the Support Vector Machine with a very low complexity constant just presents the average of the training set and thus does not offer much predictive power. However, in circumstances where the cross validation procedure has identified that no past patterns are repeated in the future and that the timeseries is just noise, the average of the training set would be identified as the best predictor.

The very high complexity constant memorizes the training set, as can be seen in the diagram where the high complexity forecast overlaps the actual demand in the training set (period 1 to 14); however, it generalizes very poorly during the testing set (period 15+). The optimal complexity constant of 0.012154154, as identified by the complexity constant optimization based on the windowed cross validation procedure (as described in the previous paragraphs), provides a forecast that represents the level of patterns learning that seems to generalize best.

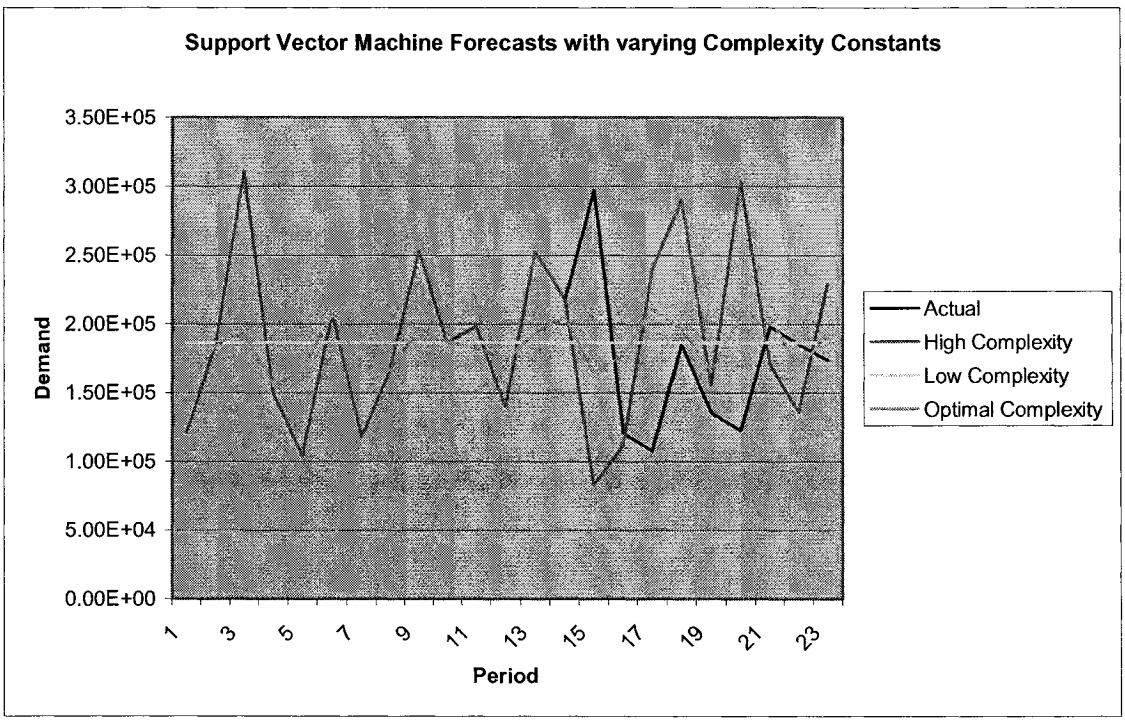

Figure 26 - SVM Forecasts with varying Complexity Constants

5.2.6 Super Wide model

As can be appreciated with the previously calculated neural-network observations-to-weights ratio, since the timeseries are very short, there are not many examples from which to learn

complex patterns. Once we have separated the data set into training, cross validation and testing sets and because we have lost periods as a result of the windowing, there are very little data left for learning. Since we have many products that probably have similar demand patterns, we will also use what we call a Super Wide model. This method takes a wide selection of timeseries from the same problem domain and combines them into one large model that will result in many training examples.

For example, in this research experiment, we consider 100 timeseries for each of the problem categories. With the Super Wide model, we use the data from all of the 100 timeseries simultaneously from one problem domain to train the model. This provides a large number of training examples and permits us to greatly increase the window size so that the models can look deep into the past data. Additionally, it could also be used to look across various other information sources that may be correlated to the demand, such as category averages or complement or substitute product demand information.

A visual representation of the Super Wide approach for simultaneously learning from multiple time-series is presented in Figure 27. Taking product 1 observation 3 (p1-o3) as an example, the demand at time period 6 (t6) is modeled as a function of the demand at time period 3 (t3), 4 (t4) and 5 (t5). This example simulates being at time period 5 (t5) and using the current demand (t5) and past 2 periods of demand (t3, t4) to predict future demand (t6). However, since this is a simulation, the learning algorithms can be taught the correction function by presenting the known demand of the simulated future period (t6). We can now take the series of observations from each product and combine them together. This is represented in the

diagram with the windowed observations of product 1 (p1) and product 2 (p2) combined together to make one large set of observations on the right side (o1 to o12).

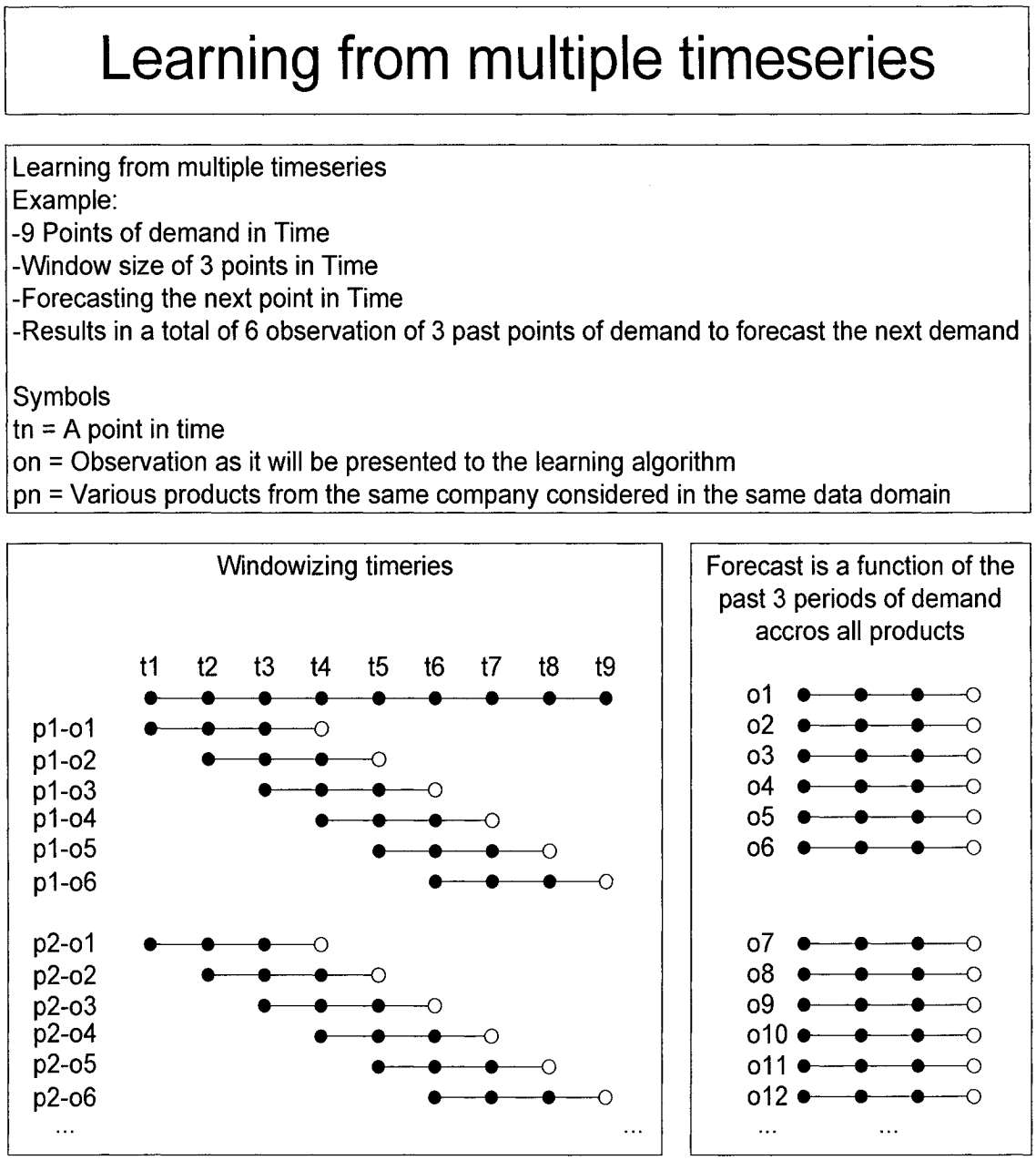

Figure 27 - Learning from multiple timeseries

An example of this model would be as follows. For the chocolate factory data set, there are 100 products and 47 periods of timeseries data. Once the training and testing set are separated, we have 38 periods of data. For this type of model, we choose a window size of 50% which is a perfect balance between modeling the demand behavior as a function of the past 50% of the data and using 50% of the data as examples. Using this large window size of 50% with the traditional timeseries model, would provide a training set of 19 examples for a window size of 19 which does not represent very much data to identify patterns that may be present in the future. However, with the Super Wide Model, we have 1900 examples for a window size of 19 which represents sufficient data to find the best forecasting patterns for the problem domain.

All of the models that learn from past demand, such as the Multiple Linear Regression, Neural Networks and Support Vector Machines will be tested also on the Super Wide models. The only exception is the Recurrent Neural Networks because the tools to do that are not yet available. Training a Recurrent Neural Network on a Super Wide model should be possible. However, it requires a reset of the recurrent connections for every product because time lagged signals between products does not make sense because the sequence of time is reset.

The Neural Network models were enlarged to 10 hidden layer neurons which in combination with the very large window provides a large input space, resulting in large network sizes compared to the patterns that were to be expected. With a window size of 50% of the training data, we have a ratio of 1 input to 1 observation. The observations are then multiplied by 100 products because of the Super Wide model format to calculate the observations to weights ratio for the chocolate manufacturer dataset.

Using the following variables:

p = number of periods in the data
n = number of products
w = window ratio
h = hidden layer neurons
b = bias (always 1)
o = number of outputs

For an Artificial Neural Network:

We can calculate the total number of weights as:

$$\text{Total Weights} = p \cdot w \cdot h + b \cdot h + h \cdot o + b \cdot o$$

Therefore for the current implementation, the number of weights will always be:

$$\text{Total Weights} = p \cdot w \cdot 10 + 1 \cdot 10 + 10 \cdot 1 + 1 \cdot 1$$

$$\text{Total Weights} = p \cdot w \cdot 10 + 21$$

And the Observations to Weights ratio is:

$$\text{Observations to Weights} = \frac{n \cdot p \cdot (1 - w)}{(p \cdot w \cdot h + b \cdot h + h \cdot o + b \cdot o)}$$

Therefore, for the chocolate manufacturer dataset, Observation to Weights ratio is:

$$\text{Observation to Weights} = \frac{100 \cdot 38 \cdot (1 - 0.05)}{(38 \cdot 0.05 \cdot 10 + 1 \cdot 10 + 10 \cdot 1 + 1 \cdot 1)}$$

ANN Observation to Weights = 90.25

For a Recurrent Neural Network:

We can calculate the total number of weights as:

$$\text{Total Weights} = p \cdot w \cdot h + b \cdot h + h \cdot h + h \cdot o + b \cdot o$$

Therefore for the current implementation, the number of weights will always be:

$$\text{Total Weights} = p \cdot w \cdot 10 + 1 \cdot 10 + 10 \cdot 10 + 10 \cdot 1 + 1 \cdot 1$$

$$\text{Total Weights} = p \cdot w \cdot 10 + 121$$

And the Observation to Weights ratio is:

$$\text{Observation to Weights} = \frac{n \cdot p \cdot (1 - w)}{(p \cdot w \cdot h + b \cdot h + h \cdot h + h \cdot o + b \cdot o)}$$

Therefore, for the chocolate manufacturer dataset, Observation to Weights ratio is:

$$\text{Observation to Weights} = \frac{100 \cdot 38 \cdot (1 - 0.05)}{(38 \cdot 0.05 \cdot 10 + 1 \cdot 10 + 10 \cdot 10 + 10 \cdot 1 + 1 \cdot 1)}$$

$$\text{RNN Observation to Weights} = 25.79$$

We can see that we now have a very high observation to weights ratio which will help us achieve much better models, especially for very noisy data.

6 Experiment Results

The performance of all of the selected models executed for the 3 datasets are presented and discussed in this section. In Table 5, we present the mean absolute errors (MAE) of all the tested forecasting techniques as applied to the Chocolate manufacturer's dataset in ascending order of error with the best techniques at the top of the list, the worst are at the bottom. The same is presented for the Toner Cartridge manufacturer's dataset (Table 6) and the Statistics Canada manufacturing datasets (Table 7). These tables also indicate the rank of each forecasting method and membership to either the control group or the treatment group of the experiment.

Chocolate Manufacturer

Table 5 - All forecasting performances for Chocolate manufacturer dataset

Rank	Cntrl./Treat.	MAE	Method	Type
1	Treatment	0.76928454	SVM CV_Window	SuperWide
2	Treatment	0.77169699	SVM CV	SuperWide
3	Control	0.77757298	MLR	SuperWide
4	Treatment	0.79976471	ANNBPCV	SuperWide
5	Control	0.82702030	ES Init	
6	Control	0.83291872	ES20	
7	Control	0.83474625	Theta ES Init	
8	Control	0.83814324	MA6	
9	Control	0.85340016	MA	
10	Control	0.86132238	ES Avg	
11	Control	0.87751655	Theta ES Average	
12	Control	0.90467127	MLR	
13	Treatment	0.92085160	ANNLMBR	SuperWide
14	Treatment	0.93065086	RNNLMBR	
15	Treatment	0.93314457	ANNLMBR	
16	Treatment	0.93353440	SVM CV	
17	Treatment	0.94270139	SVM CV_Window	
18	Treatment	0.98104892	ANNBPCV	

19	Treatment	0.99538663	RNNBPCV	
20	Control	1.01512843	ARMA	
21	Control	1.60425383	TR	
22	Control	8.19780648	TR6	

Toner Cartridge Manufacturer

Table 6 - All forecasting performances for Toner Cartridge man. dataset

Rank	Cntrl./Treat.	MAE	Method	Type
1	Treatment	0.67771156	SVM CV	SuperWide
2	Treatment	0.67810404	SVM CV_Window	SuperWide
3	Control	0.69281237	ES20	
4	Control	0.69929521	MA6	
5	Control	0.69944606	ES Init	
6	Treatment	0.70027399	SVM CV_Window	
7	Control	0.70535163	MA	
8	Control	0.70595237	MLR	SuperWide
9	Treatment	0.72214623	SVM CV	
10	Control	0.72443731	Theta ES Init	
11	Control	0.72587771	ES Avg	
12	Control	0.73581062	Theta ES Average	
13	Control	0.76767181	MLR	
14	Treatment	0.77807766	ANNLMBR	SuperWide
15	Treatment	0.80899048	RNNBPCV	
16	Treatment	0.81869933	RNNLMBR	
17	Treatment	0.81888839	ANNLMBR	
18	Treatment	0.84984560	ANNBPCV	
19	Treatment	0.88175390	ANNBPCV	SuperWide
20	Control	0.93190430	ARMA	
21	Control	1.60584233	TR	
22	Control	8.61395034	TR6	

Statistics Canada Manufacturing

Table 7 - All forecasting performances for Statistics Canada man. datasets

Rank	Cntrl./Treat.	MAE	Method	Type
1	Treatment	0.44781737	SVM CV_Window	SuperWide

2	Treatment	0.45470378	SVM CV	SuperWide
3	Control	0.49098436	MLR	
4	Treatment	0.49144177	SVM CV_Window	
5	Treatment	0.49320980	SVM CV	
6	Control	0.50517910	Theta ES Init	
7	Control	0.50547172	ES Init	
8	Control	0.50858447	ES Average	
9	Control	0.51080625	MA	
10	Control	0.51374179	Theta ES Average	
11	Control	0.53272253	MLR	SuperWide
12	Control	0.53542068	MA6	
13	Treatment	0.53553823	RNNLMBR	
14	Treatment	0.53742495	ANNLMBR	
15	Control	0.54834604	ES20	
16	Treatment	0.58718750	ANNBPCV	SuperWide
17	Treatment	0.64527015	ANNLMBR	SuperWide
18	Treatment	0.80597984	RNNBPCV	
19	Treatment	0.82375877	ANNBPCV	
20	Control	1.36616951	ARMA	
21	Control	1.99561045	TR	
22	Control	20.89770108	TR6	

From these results we can see that one of the AI approaches, the Support Vector Machine (SVM) is at the top of all three experiments by providing consistently better performance than all other techniques, however only under the Super Wide modeling approach. The results of the individual timeseries approach and the combined timeseries approach, called the Super Wide models, will be examined in detailed.

6.1 Individual timeseries models

If we ignore the Super Wide based models, we find that the results of previous research and the very large M3 competition were reproduced. That is, simple techniques outperform the more complicated and sophisticated approaches. For example, in the two primary datasets of interests, the Chocolate (Table 5) and Toner Cartridge (Table 6) manufacturer, Exponential

Smoothing has the best performance. They are at Rank 5 and Rank 3 respectively, immediately after the top Super Wide models.

This is especially true in our experiments since the data we are concerned with is very noisy and the Exponential Smoothing approach outperformed all of the other approached including all of the complex and advanced Artificial Intelligence ones by very large margins in certain cases. Noticeably, the Toner Cartridge data set was so noisy or the patterns changed so much with time that even the Exponential Smoothing with a fixed parameter of 20% outperformed (Table 6 – Rank 3) the automated one (Table 6 – Rank 5) which was optimizing the parameter for the training set. Essentially this indicates that the automated ES overfit the data, which is quite surprising.

We find the same thing with the Moving Average that is fixed to a window of 6 periods (Table 6 – Rank 4). The automatic versions have overfitting problems also and have lower performance (Table 6 – Rank 7) than setting a common parameter value (Table 6 – Rank 4). The average error of the automatic Exponential Smoothing for the two manufacturer's dataset is 0.7516 and the average for the fixed Exponential Smoothing of 20% is 0.7501 and the difference has a significance of 0.4037. The Moving Average with a window of 6 periods has a average error of 0.7561 and a mean difference significance of 0.2273 with the average of the automatic Exponential Smoothing.

However, the difference when including the Statistics Canada manufacturing data is much larger since the patterns are stronger as a result of its aggregate nature and there are more data for the automatic version to have better performance. The average error of the automatic

Exponential Smoothing for all three datasets is 0.6096 and the average for the fixed Exponential Smoothing of 20% is 0.6337 and the difference has a significance of 0.00000002159. The Moving Average with a window of 6 periods has a average error of 0.6288 and a significance of 0.0000005752 with the average of the automatic Exponential Smoothing. From this we find that for the manufacturer's datasets, there is no significant difference. Therefore, the automatic Exponential Smoothing, 20% Exponential Smoothing and the 6 period window Moving Average all provide about the same performance. However, since we are impartial for the direct manufacturing dataset and that on the larger aggregate manufacturing datasets, the Statistics Canada manufacturing survey, there was a significant different in favor of the automatic Exponential Smoothing, we would find this technique to be superior since there is added value at no loss.

In the case of the Statistics Canada dataset, the results are a little bit different; we find the MLR (Table 7 – Rank 3), SVM (Table 7 – Rank 4 and 5) and Theta (Table 7 – Rank 6) outperform Exponential Smoothing (Table 7 – Rank 7). However, because these approaches have such poor performance on the chocolate and toner cartridge manufacturer datasets and that the performance gain by these over the ES method is very small, we do not consider these results relevant. They are probably just a result of the very large amount of data (12 years) and the aggregate nature of the data which is less noisy.

It is also interesting to note that the Trend is by far the worst forecasting approach since it always ranks at the bottom of all three experiments (Rank 21 and 22) considering that this is often an informal way of planning by extrapolating that a certain trend will continue in the future. Also, ARMA and other AI approaches all did very poorly and should not have a part

in forecasting for manufacturer's in a supply chain since the data has such a low volume and is highly noisy.

6.2 Support Vector Machines using the Super Wide data

The overall best performance was obtained using Support Vector Machines in combination with the Super Wide data. Since we have previously identified that the best traditional technique is the automatic Exponential Smoothing, we can calculate the forecast error reduction provided by the above AI approach. For the chocolate manufacturer's dataset (Table 5 – Rank 2 and 5), we find a 6.70% ((0.82702030 - 0.77169699) / 0.82702030) reduction in the overall forecasting error and for the tone cartridge manufacturer dataset (Table 6 – Rank 1 and 5), we find a 3.11% ((0.69944606 - 0.67771156) / 0.69944606) reduction in the overall forecasting error. In the case of the Statistics Canada manufacturing dataset (Table 7 – Rank 2 and 7), we find a 10.00% ((0.50547172 - 0.45470378 / 0.50547172) reduction in the forecasting error as compared to automatic Exponential Smoothing. This is an average of 4.90% for our two manufacturer's dataset and an average of 6.61% for all three as compared to automatic Exponential Smoothing. The performance of the Super Wide models may even improve as more products are included beyond the limit of 100 used in our research since more examples will result in better models. To demonstrate this, we will examine in detail four major components of these results; (1) Cross Validation, (2) alternative methods, (3) t-tests and (4) sensitivity analysis.

6.2.1 Cross Validation

We tested two different Support Vector Machine cross validation based parameter optimization procedures, the windowed (time-oriented) approach and the standard approach. For the chocolate manufacturer and for the Statistics Canada datasets, the windowed cross validation was superior and for the toner cartridge manufacturer the unordered approach was better. Accordingly, we are impartial regarding the cross validation (CV) procedure.

Since the standard cross validation procedure is simpler to implement than the windowed counterpart, there can be more models tested while at the same time using more data for each CV model. Furthermore, since the error curves seemed more stable (have lower variation and clearer concave shape), we recommend the standard CV procedure over the windowed one. It is interesting to further examine how the cross validation based parameter selection behaves. It is this key feature in combination with the guaranteed optimality of the SVM which makes it possible to determine the best level of complexity. The cross validation error curves for the range of complexity constants is presented in Figure 28 for the chocolate manufacturer's dataset, Figure 29 for the toner cartridge manufacturer's dataset and Figure 30 for Statistics Canada manufacturing dataset. In both Figure 28 and Figure 29 there is a clear concave pattern which indicates that a complexity constant, that generalizes well, is identified without ambiguity. The optimal complexity is more difficult to identify in Figure 30 because as the complexity increases, the error stays relatively low and stable. This is probably a result of the larger amount of data and less noise and so there is a range of complexity which may generalize well. In all three figures, there is a clear distinction between complexity levels that generalize better than others thus permitting the selection of a complexity level. By contrast, if

these figures presented error lines that randomly moved up and down as the complexity constant varied, this would indicate that the cross validation procedure was not providing any value.

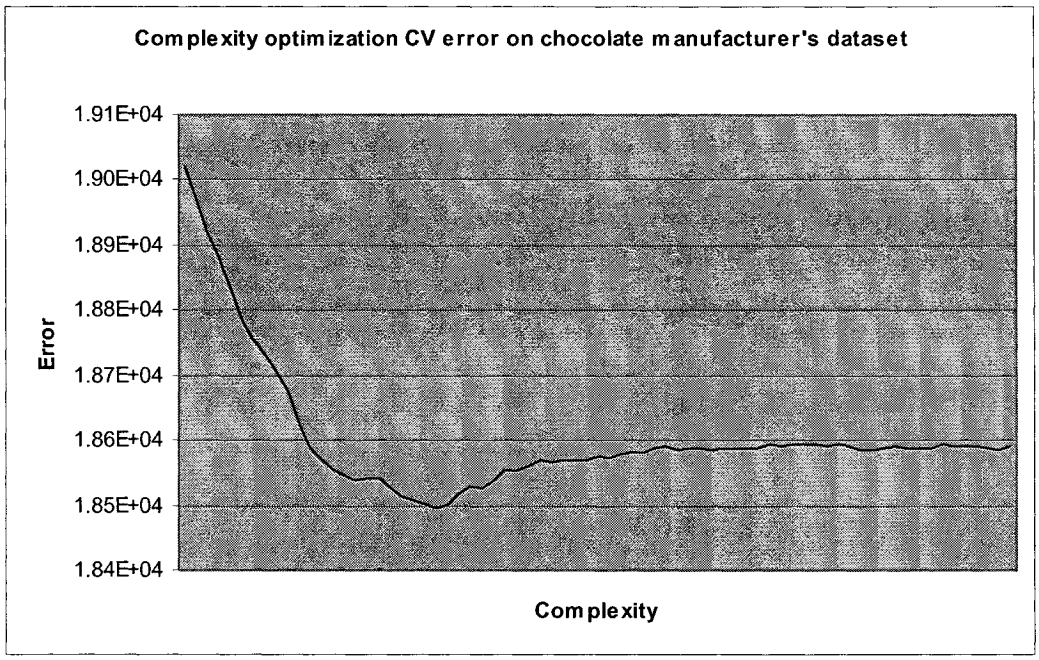

Figure 28 - Complexity optimization CV error on chocolate man. dataset

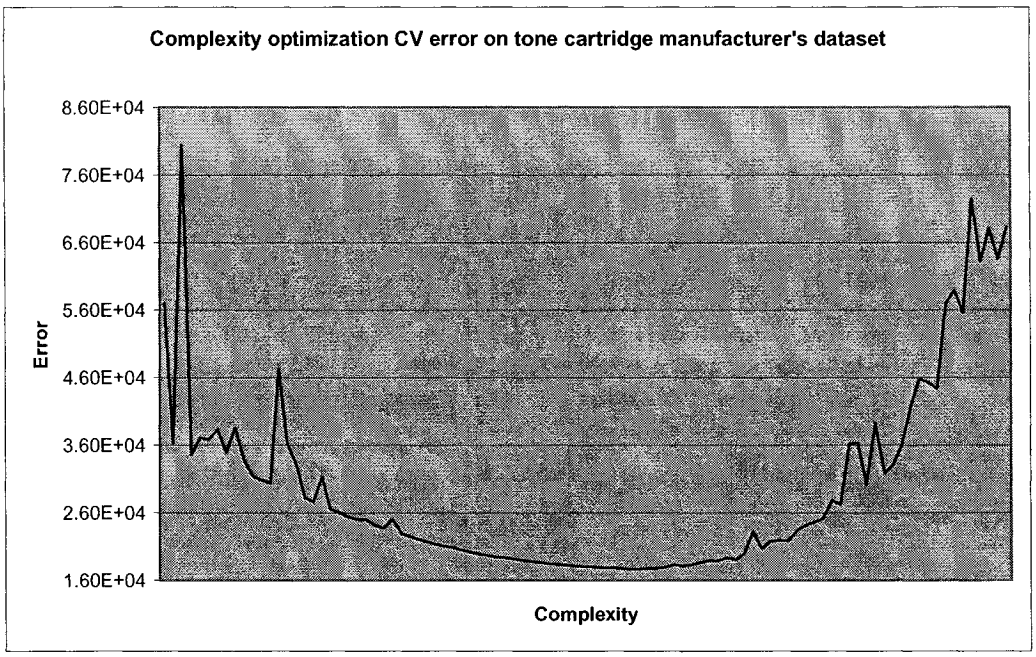

Figure 29 - Complexity optimization CV error on tone cartridge man. dataset

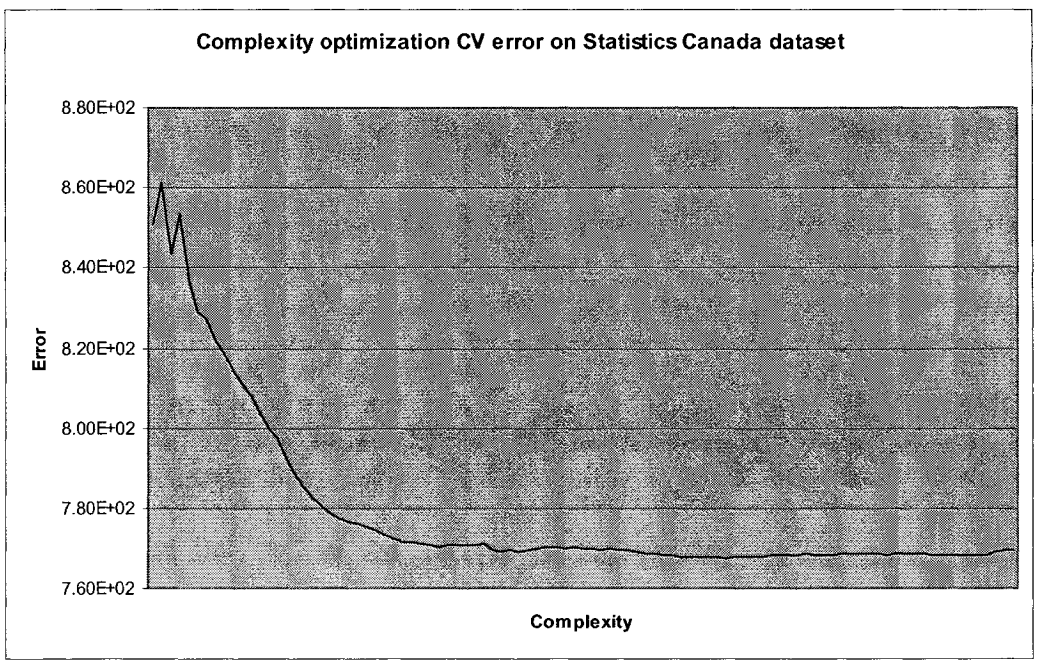

Figure 30 - Complexity optimization CV error on Statistics Canada dataset

6.2.2 Alternatives

In the case of the chocolate manufacturer's dataset, we find that the next best performing algorithms that are better than the Exponential Smoothing are the Super Wide Multiple Linear Regression (MLR) and the Super Wide Artificial Neural Networks (ANN) with Cross Validation based early stopping. In analyzing the ANN's performance, we can see that, even though its performance is strong on the chocolate manufacturer's dataset (Table 5 – Rank 4), it has extremely poor performance on both the toner cartridge manufacturer dataset (Table 6 – Rank 19) and on the Statistics Canada manufacturing dataset (Table 7 – Rank 16). As a result, the Super Wide Artificial Neural Networks with Cross Validation based early stopping method must be disregarded as a potential alternative.

The Super Wide Multiple Linear Regression (MLR) is much closer with an average error across the two manufacturer's dataset of 0.7353 compared to 0.7516 (significance 0.0274) for the automatic Exponential Smoothing (ES) and compared to 0.7154 (significance 0.0000000264) for the Super Wide SVM. The averages across all three datasets are 0.6184 for Super Wide MLR compared to 0.6096 (significance 0.1066) for ES and compared to 0.5651 (significance 2.06E-31) for Super Wide SVM.

From these results, we find that depending on what p-value is considered significant, the MLR performs slightly better than ES for the two manufacturer's datasets, but performs worse than ES when considering all three datasets. So we cannot clearly state whether a linear version of the Super Wide model is better than the traditional and simple ES. However, we do find that results using Super Wide Support Vector Machine model are significantly better than the

Multiple Linear Regression, its linear counterpart, which would indicate that the performance increase identified is not isolated to only the Super Wide model. Additionally, since in almost all of the cases both non-linear and linear alternative techniques are much worse than the Exponential Smoothing, the high quality modeling performed by the Super Wide Support Vector Machine seems to be a result of the combination of the Super Wide data and the Support Vector Machine and not a result of either method applied separately.

6.2.3 Average Performance

Earlier we stated our hypothesis 1, that Artificial Intelligence based techniques used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better average performance than traditional techniques.

To test this hypothesis, we examine the difference between the average error of traditional forecasting techniques and AI based techniques. This is important for those who are concerned with the expected error for randomly selecting a traditional forecasting technique compared to randomly selecting an AI technique. By taking the average error of the control and treatment group we can evaluate if AI in general presents a better solution.

Before proceeding, we must note that the Trend technique provides extremely poor forecasting and the error measurements are extreme outliers. However, there are not outliers in the sense of being a measurement error, they are correct values and must be retained in the average calculation. For example, if a practitioner was to randomly choose a traditional technique for each of his company's 1000 products, the extremely poor forecasting quality of the trend would affect his forecasts. Thus affecting his inventory and consequently the

company's financial performance, regardless of whether the Trend is an outlier or not. The same can be applied to the average performance of many companies each randomly selecting one traditional forecasting technique to use. Additionally, the Trend forecasting techniques must be retained in the experiment because it is representative of what practitioners use since according to a survey by Jain (2004c), averages and simple trend are used 65% of the time.

Using the results of the experiments on chocolate (Table 5) and toner cartridge (Table 6) manufacturer, the average error in the control group is 1.5014 and the average error in the treatment group is 0.8356. It was not feasible to take all of the error points of the test sets for each forecasting technique and compare those of the control group with the treatment group since there would be 52800 observations for the control group and 44000 observations for the treatment group. However, considering the large difference between the averages and the large number of observations, the t-test would have a very high significance value and we consider this hypothesis as supported.

Accordingly, without any specific information on which traditional and AI techniques are best or if there is a lack of experience and knowledge related to traditional and AI techniques, one can be safe in choosing a random AI solution to provide the lowest expected error.

6.2.4 Rank Performance

Earlier we stated our hypothesis 2, that Artificial Intelligence based techniques used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better ranked performance than traditional techniques.

Since the first hypothesis is subject to the impact of Trend's very poor performance, a second hypothesis presents a rank based point of view which is not affected by the large error of the Trend forecasting. By taking the average rank of the control and treatment group we can evaluate if AI in general presents a better solution. Using the results of the experiments on chocolate (Table 5) and toner cartridge (Table 6) manufacturer, the average rank in the control group is 11.25 and the average rank in the treatment group is 11.8. Therefore, on average, artificial intelligence techniques ranked worse than the traditional ones. Since the averages present results that contradict the hypothesis, the hypothesis is rejected and a t-test is not required. If the Statistics Canada data is included, the averages rank of both the control and the treatment group are 11.5, thus also rejecting the hypothesis.

Accordingly, without any specific information on which traditional and AI techniques are best or if there is a lack of experience and knowledge related to traditional and AI techniques, one cannot just select any AI method and generally expect it to perform better than a random traditional technique.

However, assuming rational practitioners, there would be an attempt to identify the technique that provides the best forecasts and only implement this one. They would not be concerned with the averages or average rank of either group since random selection would only be

considered if practitioners considered all techniques equal, which they clearly are not. For example, we find that the Trend models perform very poorly, so even if the group of artificial intelligence based forecasting algorithms was not shown to be better than the traditional group, a practitioner would not want to use the Trend models even if they are part of the traditional group. To address the specific issue of the best method, the next test examines the best methods from the control and treatment group.

6.2.5 Comparison of the Best AI and traditional technique

Earlier we stated our hypothesis 1, that the best Artificial Intelligence based technique used at the end of the supply chain to forecast distorted customer demand as experienced by a manufacturer will have better performance than the best traditional technique.

To statistically compare the performance between the best identified artificial intelligence based forecasting techniques and the best traditional forecasting technique for forecasting manufacturers' demand, we will perform a t-test. All of the errors have been converted to normalized absolute errors which provide us with a series of forecast for both the test period of 20% of the dataset and all of the products which can be compared to identify if there is a statistical difference. The total forecast points verified for the chocolate manufacturer's dataset is 900 and the total for the toner cartridge manufacturer is 1300. Together this is a total of 2200 datapoints which are used in the t-test to compare the errors of the Super Wide Support Vector Machine with standard Cross Validation based parameter selection with the Exponential Smoothing approach. The result of the one-tailed t-test is a difference between the Support Vector Machine error average of 0.7154 and the automatic Exponential

Smoothing error average of 0.7516 which has a significance of 0.0000058869. If we also include the Statistics Canada dataset errors, which has 3000 testing set datapoints, the total number of datapoints used in the t-test increases to 5200. Across the three datasets the Support Vector Machine error average is 0.5651 and the automatic Exponential Smoothing error average is 0.6096 for a difference significance of 6.08E-13. So we can conclude that the best AI approach is significantly better than the best traditional approach.

6.2.6 Sensitivity Analysis

The usage of the super wide data for the Support Vector Machine modeling does permit analyzing much more data simultaneously and thus permits us to set a very high historical window size while still having a very large dataset to learn from. As a result of this, we set the historical window size to 50% of the history. However, one may ask, "how do we know that 50% is the correct setting and does this setting have an impact on the performance of the model?" Although a historical window size of 50% seems natural, we re-execute the Super Wide Support Vector Machine models for a window size of 40% and one of 60% to evaluate the impact of this choice. The Mean Absolute Errors for the Super Wide Support Vector Machine with parameter optimization via standard Cross Validation for a historical window size of 40%, 50% and 60% are presented in Table 8.

Table 8 - Sensitivity analysis of window size

DataSet	Window	MAE
Chocolate	0.40	0.78060253
Chocolate	0.50	0.77169699
Chocolate	0.60	0.77028325
TonerCartride	0.40	0.67698769
TonerCartride	0.50	0.67771156

TonerCartride	0.60	0.68461909
StatsCan	0.40	0.39026298
StatsCan	0.60	0.44143398
StatsCan	0.50	0.45470378

From these results we find that for the chocolate manufacturer's dataset the error decreases when the window size increases, however, for the toner cartridge dataset, the error increases with the window size increase. The Statistics Canada manufacturing dataset shows mixed results with no trend. So we do not see a trend that would indicate that a smaller or larger window size would have an impact on the performance. Additionally, the performance difference between such large window size change (10%) result in very small differences in performance and all of these performances remain better than the identified next contender which is the Exponential Smoothing. Thus we find that the Super Wide Support Vector Machine with parameter optimization via standard Cross Validation is relatively insensitive to the window size and that a window size of 50% seems to be a good choice.

7 Conclusion and Discussion

In this research we have found some important answers to issues that relate to a manufacturer in a supply chain, who receives an extremely distorted demand signal which has a high noise to pattern ratio. Although there are several forecasting algorithms available to a practitioner, there are very few objective and reproducible guidelines to which method should be used and how. In this research, we have shown empirically that the best traditional method for a manufacturer is the automatic Exponential Smoothing with the first value as the series as the initial value and not the average. We have also found that all of the more advanced techniques including the Artificial Intelligence techniques have extremely poor performance as a result of the noisy nature of the data and the small amount of examples (time periods / months) for one product. None of the Artificial Intelligence techniques can reliably outperform the best traditional counterpart, which is the Exponential Smoothing, when learning and forecasting single time-series. Thus, they are not recommended as forecasting techniques for noisy demand at the manufacturer's end of the supply chain.

However, this research also identifies the usefulness of combining the data of multiple products, in what we call a Super Wide model, in conjunction with a relatively new technique, the Support Vector Machine and a cross validation based parameter optimization. The domain specific empirical results show that this approach is superior to the next best counterpart which is the Exponential Smoothing. The error reduction found range from 3.11% to 10% which can result in a large financial savings for a company depending on the cost related to inventory errors. This assumes that the company is already using Exponential

Smoothing, the optimal forecasting method for manufacturers. If they are not, the performance gains would be much greater.

We also feel that as the number of products added to the combined timeseries model (Super Wide approach) increases, the performance will also probably increase further since there will be more data to learn from. As any business decision, the use of the technology presented here should be based on a cost-benefit analysis of the benefits of implementing such technology weighed against its cost. If this approach is viable, in the long run it may be integrated into enterprise resource planning systems for automated and interaction free forecasting.

The potential applications of the results found in this research go further than just use by organizations that do not perform collaborative forecasting. Collaborative planning provides a great opportunity for reducing the total supply chain cost and help the manufacturer either deal with or reduce the large amount of distortion and noise in the demand patterns. However, collaborative forecasting is still required to extrapolate patterns into the future and anytime that data is added to forecasting models, we increase the input dimensionality of the model without increasing the number of examples to learn from. The Super Wide approach in combination with the SVM may provide an opportunity for merging more data into forecasting models while still having a large enough dataset to learn these complex patterns. This additional data may include economic indicators, market indicators, collaborative information sources and product group averages for example. Thus the findings in this research may lead to a much wider application for extracting patterns from time dependent data in today's organizations.

7.1 Generalization

The results of this research should generalize very well to the specific problem domain which is the manufacturer at the end of a supply chain since we have actual data from a large number of products from two North American manufacturers. Additionally, the results have been verified again Statistics Canada manufacturing survey, which provides dataset from across all Canadian manufacturing industries, thus increasing the generalizability of the results.

7.2 Application

One important point to note is that Support Vector Machines are computationally intensive and the cross validation based complexity parameter optimization procedure results in running a large amount of support vector machines depending on the precision of the complexity search. In our research we have been able to reduce the search space because we know the nature of the data and the realistic search range and we have also reduced the data set size to 100 products for each manufacturer. The longest running models in this research took over 3 days of processing on a modern computer in 2005. It is reasonable to extrapolate that application of these techniques today would result in models that have execution times of several weeks. There are many optimizations that could be performed to reduce the processing time such as parallelization which is trivial for a Cross Validation procedure and reduction of the complexity term search precision.

Realistically, the technology presented in this research may take 3-5 years before it would be integrated into existing systems and software assuming that it is an acceptable and superior approach. We could guess that by that time, further optimizations to the Support Vector

Machine algorithms and the increase in processing power would reduce the processing time by 10-fold before parallelization which would mean that a model would take a couple days to process. If the processing is parallelized across an organization's computers, the processing time should not be more than one night. Once the models have been completed, they can be used for forecasting with relatively little processing time.

There is also research into hardware based Support Vector Machine implementations. One such project is the Kerneltron which provide performance increases by a factor of 100 to 10,000 (Genov and Cauwenberghs 2001; Genov and Cauwenberghs 2003; Genov, Chakrabartty et al. 2003), thus increasing the probability that such large SVM applications are feasible in the medium term future.

References

Anonymous (2005). "Inventory Carrying Costs." The Controller's Report(4): 5.

Assimakopoulos, V. and K. Nikolpoulos (2000). "The theta model: A decomposition approach to forecasting." International Journal of Forecasting **16**(4): 521.

Box, G., G. M. Jenkins, et al. (1994). Time Series Analysis: Forecasting and Control. Englewood Cliffs, NJ: Prentice Hall.

Chandra, C. and J. Grabis (2005). "Application of multi-steps forecasting for restraining the bullwhip effect and improving inventory performance under autoregressive demand." European Journal of Operational Research **166**(2): 337.

Chatfield, C. (2001). "Is 'bigger' necessarily 'better'?" International Journal of Forecasting **17**(4): 547-549.

Cox, A., J. Sanderson, et al. (2001). "Supply chains and power regimes: Toward an analytic framework for managing extended networks of buyer and supplier relationships." Journal of Supply Chain Management **37**(2): 28.

Davis, E. W. and R. E. Spekman (2004). The extended enterprise: gaining competitive advantage through collaborative supply chains. Upper Saddle River, NJ : FT Prentice Hall.

de Figueiredo, R. J. P. (1980). "Implications and applications of Kolmogorov's superposition theorem." IEEE Transactions on Automatic Control **25**(6): 1227–1231.

Dejonckheere, J., S. M. Disney, et al. (2003). "Measuring and avioiding the bullwhip effect: A control theoretic approach." European Journal of Operational Research **147**(3): 567.

Dorffner, G. (1996). "Neural networks for time series processing." Neural Network World **96**(4): 447-468.

Elman, J. L. (1990). "Finding structure in time." Cognitive Science **14**(2): 179-211.

Foresee, F. D. and M. T. Hagan (1997). "Gauss-Newton approximation to Bayesian regularization." Proceedings of the 1997 International Joint Conference on Neural Networks: 1930-1935.

Forrester, J. (1961). Industrial Dynamics. Cambridge, MA: Productivity Press.

Genov, R. and G. Cauwenberghs (2001). "Charge-Mode Parallel Architecture for Matrix-Vector Multiplication." IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing **48**(10): 930-936.

Genov, R. and G. Cauwenberghs (2003). "Kerneltron: Support Vector "Machine" in Silicon." IEEE Transactions on Neural Networks 14(5): 1426-1434.

Genov, R., S. Chakrabartty, et al. (2003). "Silicon Support Vector Machine with On-Line Learning." International Journal of Pattern Recognition and Artificial Intelligence 17(3): 385-404.

Giles, C. L., S. Lawrence, et al. (2001). "Noisy Time Series Prediction using Recurrent Neural Networks and Grammatical Inference." Machine Learning 44: 161-184.

Gunasekaran, A. and E. W. T. Ngai (2004). "Information systems in supply chain integration and management." European Journal of Operational Research 159(2): 269.

Hagan, M. T., H. B. Demuth, et al. (1996). Neural Network Design. Boston, MA: PWS Publishing.

Hagan, M. T. and M. Menhaj (1994). "Training feedforward networks with the Marquardt algorithm." IEEE Transactions on Neural Networks 5(6): 989–993.

Heikkila, J. (2002). "From supply to demand chain management: Efficiency and customer satisfaction." Journal of Operations Management 20(6): 747.

Herbrich, R., M. Keilbach, et al. (1999). Neural networks in economics: Background, applications and new developments. Boston, MA: Kluwer Academics.

Hornik, K. (1991). "Approximation capabilities of multilayer feedforward networks." Neural Networks 4(2): 251-257.

Jain, C. L. (2004a). "BENCHMARKING FORECASTING ERROR." The Journal of Business Forecasting Methods & Systems 23(3): 8.

Jain, C. L. (2004b). "BENCHMARKING FORECASTING SOFTWARE AND SYSTEMS." The Journal of Business Forecasting Methods & Systems 23(3): 13.

Jain, C. L. (2004c). "BUSINESS FORECASTING PRACTICES IN 2003." The Journal of Business Forecasting Methods & Systems 23(3): 2.

Joachims, T. (1999). "Making large-Scale SVM Learning Practical." Advances in Kernel Methods - Support Vector Learning.

John Galt Solutions, Inc. (2005). Forecast Expert Toolkit Chicago, IL: John Galt Solutions Inc.

Kimbrough, S. O., D. J. Wu, et al. (2002). "Computers play the beer game: Can artificial agents manage supply chains?" Decision Support Systems 33(3): 323.

Lambert, D. M. and B. J. Lalonde (1976). "INVENTORY CARRYING COSTS." Management Accounting **58**(2): 31.

Landt, F. W. (1997). Stock Price Prediction using Neural Networks. Computer Science. Leiden, Netherlands, Leiden University. **Master:** 62.

Lee, H. L., V. Padmanabhan, et al. (1997a). "The Bullwhip Effect in Supply Chains." Sloan Management Review **38**(3): 93.

Lee, H. L., V. Padmanabhan, et al. (1997b). "Information distortion in a supply chain: The bullwhip effect." Management Science **43**(4): 546.

MacKay, D. J. C. (1992). "Bayesian Interpolation." Neural Computation **4**(3): 415-447.

Makridakis, S., A. Andersen, et al. (1982). "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition." Journal of Forecasting (pre-1986) **1**(2): 111.

Makridakis, S., C. Chatfield, et al. (1993). "The M2-Competition: A real-time judgmentally based forecasting study." International Journal of Forecasting **9**(1): 5.

Makridakis, S. and M. Hibon (1979). "Accuracy of forecasting: an empirical investigation (with discussion)." Journal of the Royal Statistical Society. Series A (General) **142**(2): 97-145.

Makridakis, S. and M. Hibon (2000). "The M3-Competition: Results, conclusions and implications." International Journal of Forecasting **16**(4): 451.

Marquardt, D. W. (1963). "An algorithm for least-squares estimation of nonlinear parameters." SIAM Journal of Applied Mathematics **11**: 431-441.

MathWorks, Inc. (2005a). Financial Time Series Toolbox for Use with MATLAB. Natick, MA: MathWorks Inc.

MathWorks, Inc. (2005b). GARCH Toolbox for Use with MATLAB. Natick, MA: MathWorks Inc.

MathWorks, Inc. (2005c). Getting Started with MATLAB. Natick, MA: MathWorks Inc.

MathWorks, Inc. (2005d). Neural Network Toolbox for Use with MATLAB. Natick, MA: MathWorks Inc.

MathWorks, Inc. (2005e). Optimization Toolbox for Use with MATLAB. Natick, MA: MathWorks Inc.

Microsoft, Corp. (2005). Excel Statistics Analysis ToolPack Redmond, WA: Microsoft Corp.

Mukherjee, S., E. Osuna, et al. (1997). Nonlinear prediction of chaotic time series using support vector machines. {IEEE} Workshop on Neural Networks for Signal Processing {VII}, Ameila Island, FL, USA, IEEE Press.

NeuroSolutions, Inc. (2005). NeuroSolutions for Excel. Gainesville, FL: NeuroSolutions Inc.

Ord, K. (2001). "Commentaries on the M3-Competition: An introduction, some comments and a scoreboard." International Journal of Forecasting **17**(4): 537-584.

Premkumar, G. P. (2000). "Interorganization systems and supply chain management: An information processing perspective." Information Systems Management **17**(3): 56.

Proctor and Gamble, Co (2005). 2004 Annual Report. Cincinnati, Proctor and Gamble Co.

Raghunathan, S. (1999). "Interorganizational collaborative forecasting and replenishment systems and supply chain implications." Decision Sciences **30**(4): 1053.

Rumelhart, D. E., G. E. Hinton, et al. (1986). "Learning internal representations by error propagation." Parallel Distributed Processing **1**: 318-362.

Rüping, S. (2005). mySVM-Manual, Universitat Dortmund, Lehrstuhl Informatik 8.

Rüping, S. and K. Morik (2003). Support Vector Machines and Learning about Time. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong.

Russell, V. L. (2001). "Some practical guidelines for effective sample size determination." The American Statistician **55**(3): 187.

SAP, AG (2005). SAP R/3 Enterprise. Walldorf, SAP AG.

SAS Institute, Inc. (2005). SAS High-Performance Forecasting Cary, NC: SAS Institute Inc.

Statistics Canada (2005). Monthly Survey of Manufacturing (Code 2101), Statistics Canada.

Stitt, B. (2004). "DEMAND PLANNING: PUSHING THE REST OF THE COMPANY TO DRIVE RESULTS." The Journal of Business Forecasting Methods & Systems **23**(2): 2.

Tan, K. C. (2001). "A framework of supply chain management literature." European Journal of Purchasing & Supply Management **7**(1): 39-48.

Thonemann, U. W. (2002). "Improving supply-chain performance by sharing advance demand information." European Journal of Operational Research **142**(1): 81.

Vakharia, A. J. (2002). "e-business and supply chain management." Decision Sciences **33**(4): 495.

Vapnik, V., S. Golowich, et al. (1997). "Support vector method for function approximation, regression estimation, and signal processing." Advances in Neural Information Systems **9**: 281-287.

Vapnik, V. N. (1995). The nature of statistical learning theory. New York, Springer-Verlag.

Werbos, P. J. (1990). "Backpropagation through time: what it does and how to do it." Proceedings of the IEEE **78**(10): 1550 - 1560.

Wisner, J. D. and L. L. Stanley (1994). "Forecasting practices in purchasing." International Journal of Purchasing and Materials Management **30**(1): 22.

Yusuf, Y. Y., A. Gunasekaran, et al. (2004). "Agile supply chain capabilities: Determinants of competitive objectives." European Journal of Operational Research **159**(2): 379.

Zhao, X., J. Xie, et al. (2002). "The impact of forecast errors on early order commitment in a supply chain." Decision Sciences **33**(2): 251.

Appendix I – Selected Cat. from Statistics Canada Manufacturing Survey

Category
Men's and boys' cut and sew shirt manufacturing
Glass manufacturing
Narrow fabric mills and Schiffli machine embroidery
Metal window and door manufacturing
Fabric coating
Office furniture (including fixtures) manufacturing
Motor vehicle metal stamping
Resin and synthetic rubber manufacturing
Other women's and girls' cut and sew clothing manufacturing
Women's and girls' cut and sew blouse and shirt manufacturing
Cold-rolled steel shape manufacturing
Automobile and light-duty motor vehicle manufacturing
Prefabricated metal building and component manufacturing
Mattress manufacturing
Fertilizer manufacturing
Power boiler and heat exchanger manufacturing
Stationery product manufacturing
Paint and coating manufacturing
Other industrial machinery manufacturing
Rubber and plastic hose and belting manufacturing
Hosiery and sock mills
Commercial and service industry machinery manufacturing
Particle board and fibreboard mills
Non-ferrous metal foundries
All other converted paper product manufacturing
Wood window and door manufacturing
Boat building
Wineries
All other general-purpose machinery manufacturing
Pulp mills
Non-chocolate confectionery manufacturing
Plastic bottle manufacturing
Men's and boys' cut and sew trouser, slack and jean manufacturing
Copper rolling, drawing, extruding and alloying
Audio and video equipment manufacturing
Flour milling and malt manufacturing
Dairy product manufacturing
Railroad rolling stock manufacturing

Forging and stamping
Institutional furniture manufacturing
Ready-mix concrete manufacturing
Support activities for printing
Industrial gas manufacturing
Household furniture (except wood and upholstered) manufacturing
Other concrete product manufacturing
Non-ferrous metal (except copper and aluminum) rolling, drawing, extruding and alloying
Motor vehicle transmission and power train parts manufacturing
Spring and wire product manufacturing
Synthetic dye and pigment manufacturing
Sawmill and woodworking machinery manufacturing
Textile and fabric finishing
Jewellery and silverware manufacturing
Motor and generator manufacturing
Concrete reinforcing bar manufacturing
Corrugated and solid fibre box manufacturing
Motor vehicle transmission and power train parts manufacturing
Pump and compressor manufacturing
Other men's and boys' cut and sew clothing manufacturing
Women's and girls' cut and sew lingerie, loungewear and nightwear manufacturing
Nonwoven fabric mills
Other paperboard container manufacturing
Carpet and rug mills
Aerospace product and parts manufacturing
Poultry processing
Explosives manufacturing
Cut and sew clothing contracting
Motor home, travel trailer and camper manufacturing
Folding paperboard box manufacturing
Pharmaceutical and medicine manufacturing
Office supplies (except paper) manufacturing
Battery manufacturing
Small electrical appliance manufacturing
Construction machinery manufacturing
Plastics pipe, pipe fitting, and unlaminated profile shape manufacturing
Other transportation equipment manufacturing
Petroleum refineries
Motor vehicle gasoline engine and engine parts manufacturing
Starch and vegetable fat and oil manufacturing
Knit fabric mills
Heavy-duty truck manufacturing
Artificial and synthetic fibres and filaments manufacturing
Paper (except newsprint) mills
Women's and girls' cut and sew suit, coat, tailored jacket and skirt manufacturing

Radio and television broadcasting and wireless communications equipment manufacturing
Soap and cleaning compound manufacturing
Power, distribution and specialty transformers manufacturing
Shingle and shake mills
Adhesive manufacturing
Motor vehicle plastic parts manufacturing
Wood preservation
Toilet preparation manufacturing
Other basic inorganic chemical manufacturing
Wood kitchen cabinet and counter top manufacturing
Wiring device manufacturing
Sign manufacturing
Steel foundries
Non-ferrous metal (except aluminum) smelting and refining
Semiconductor and other electronic component manufacturing
Urethane and other foam product (except polystyrene) manufacturing
Sawmills (except shingle and shake mills)