Table of Contents

1 Introduction

Accurate sales forecasting is of paramount importance to many business activities such as Supply chain management, Inventory management and long term sales strategies. Organizations investing in forecasting best practices gain a competitive advantage over rival organizations. For demand sales, times series based forecasting methods have been popular and well researched.

The biggest challenge in forecasting a sales series, is to identify the best forecasting model for a given times series. This is an extremely difficult task, since sales series are too complex for one model to work best at all times. A realistic situation that could occur is, one choosing the best one is not really an option. Thus researchers have long since been advocating the use of a "Combination Forecast".

In combination forecasting, instead of trying to choose the single best method, there is an attempt to identify a group of methods which would in conjunction, help to improve group of methods which would in conjunction, help to improve group of methods which would in conjunction, help to improve group of methods which would in conjunction, help to improve forecast accuracy.

Another approach that has met with much success is to first decompose a time series into its "natural components" and then do the forecasting. Recent work in this area has effectively used a decomposition followed by combination for successful demand forecasting. For sales data, the natural components include trend, seasonality, and an irregular component. Each component is considered as an individual series and a battery of models is utilized to forecast each component series. Each decomposed series is forecasted using a set of models. These are recombined to generate a large number of combination forecast values per point a point forecast based on such a large number of estimates poses both an opportunity and a challenge.

Data Mining provides us with a solution to this problem. Over the last fifteen years or so, a rich store of techniques has been compiled to mine large data repositories for valuable patterns, associations, correlations, trends, etc. Recent work in data mining for times series forecasting, uses frequent pattern mining to "learn" a group of forecasters that tend to have superior forecasting performance for a part of the series (during the training phase). These forecasters are further used, to forecast the rest of the series.

This work has further been extended by using a fine grained approach which identifies a set of Seasonal, Trend and Irregular experts for a given series.

A combination of such methods has allowed a very fine grained, yet relevant forecasts to be calculated using the S,T,I experts and groups of well-preforming forecasters.

The complimentary problem of identifying a poorly-performing set of forecasters and then eliminating these during the testing phase has also been tested Both these algorithms have led to increased accuracy over the traditional standard model for demand sales, the Holt–Winter method.

The decomposition approach is used in conjunction with tens of forecasters per component. This results in thousands of combination forecast values per point to be forecast. Computing a point forecast based on such a large number of estimates poses both an opportunity and a challenge.

Over the last fifteen years or so, a rich store of techniques has been compiled to mine large data repositories for valuable patterns, associations, correlations, trends, etc.
These methods allow analysts to produce multidimensional forecasts with relation to volumes, correlation among products, and relevant trends

Huge retail chains have thousands of items in their inventory. They store a huge volume of data to keep track of business trends. Ultimately this data can be useful only if it is subjected to data mining and allied techniques.

The task of forecasting the sales of each and every product using the above described frequent pattern mining based algorithm, is computationally expensive and practically infeasible. Our project proposes an innovative way to overcome the above problem. We identify clusters within the inventory consisting of items which have the same sales patterns. These clusters are identified by just analyzing the nature of the sales times series of each individual item.

Our methods of classifying series additionally employ a novel distance measure which has been named as Sales Pattern Distance. This measure quantifies similarity or the lack of it between two series by identifying whether or not the series follow a similar pattern

Using the earlier mentioned algorithms we identify the good and bad set of Trend and Irregular experts for one representative item of each cluster. These set of experts are used to forecast sales for all items of the cluster. Then we calculate MAPE for these forecasts, which serve as a metric of accuracy for the forecasted data.

We propose to show that our algorithm in addition to being efficient also reduces the forecasting error considerably in comparison to the standard Holt Winter approach.

2 Literature Survey

In this chapter we provide a brief review of the literature that is relevant to this research.

2.1 Classification of forecasting methods

Forecasting methods can be classified into two broad categories: methods that derive primarily from judgmental sources (Qualitative) versus those from statistical sources (Quantitative). These methods and their relationships are shown in the flow chart in Figure 1.1. Makridakis, Wheelwright and Hyndman (1998) provide details on how to apply many of these methods.

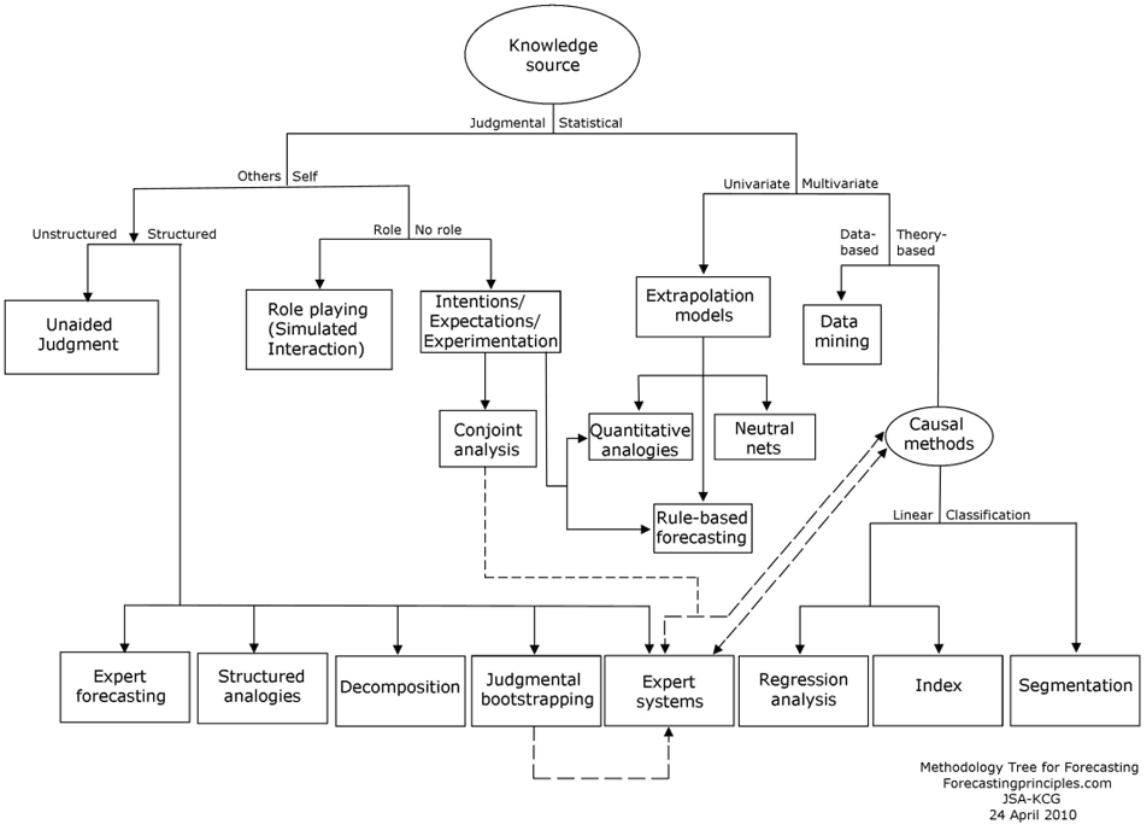


Figure 2-1 Forecasting Methodology Tree

2.1.1 Qualitative methods –

These methods do not conform to any formal mathematical model. These methods are called judgmental, and they mainly use human expertise and intuition rather than any formal technique. The inputs required depend on the specific method and are mainly the product of judgment and accumulated knowledge.These methods are normally used in cases where the data available is not

thought to be representative of the future (long-term forecasting). These methods require inputs from a set of highly specialised human 'experts' in the field. It is widely reported that (Smith et al. 1996) reported that as an enterprise becomes mature the forecasting process moves from a qualitative one to a quantitative process. On the other hand, companies in their initial stages are more people oriented and highly subjective and qualitative in their forecasting processes. Table 1 provides a brief description of the more popular qualitative forecast techniques.

Name	Description
Grass Roots	Derives a forecast by compiling input from those at the end of the hierarchy who are involved in the forecasting procedure.
Market Research	Sets out to collect data in a variety of ways (surveys, interviews etc.) to test hypotheses about the market. It is usually used for long term forecasting
Panel Consensus	Free open exchange at meetings. The idea is that the discussion within the group will produce better forecasts that any individual.
Historical Analogy	Ties what is being forecast to a similar item. Forecasts are then based on the historical pattern of the similar item.
Delphi Method	Group of experts responds to questionnaire. A moderator compiles results and formulates a new questionnaire that is submitted to the group. This phase is repeated for a number of times until a forecast emerges.
Sales Force Composite	Members of the sales force submit their estimates based on their knowledge acquired by the daily interaction with customers
Customer Survey	Customer surveys can be used to signal future trends and shifting preference patterns
Cross-Impact Analysis	It is an approach used often in conjunction with Delphi method. The method attempts to assess the interdependence of uncertain future events. By doing so, it helps forecasters to understand what developments are likely to influence the area of interest so better forecasts can be generated.
Scenario Writing	It is based on writing of scenarios, in which views of possible futures are verbally exposited.
Economic Indicators	Economic indicators are tracked across a time series. The description of the series provides a basis to construct a judgmental forecast.

Table 2-1 (Source: Chase et al.,2006; Nahmias, 2005; Newbold and Bos, 1994)

2.1.2 Quantitative Methods:

These can be applied when three conditions exist: Information about the past is available; information can be quantified in the form of numerical data; it can be assumed that some aspects of the past pattern will continue into the future. Quantitative techniques consist of two categories: Explanatory (causal) models and time series models.

Explanatory methods

These methods work under the assumption that it is possible to identify the underlying factors that might influence the variable that is being forecast. They thus form the final forecast as a function of these independent variables (factors).

For e.g. $Sales\ of\ shoes\ =\ function\ (time\ of\ year, season, month\ \ldots..)$

Extensive theory, vast prior research, and specialized expert domain knowledge are needed to gain information about relationships between explanatory variables and the variable to be forecast. Causal models are most useful when (1) strong causal relationships exist, (2) the directions of the relationships are known, (3) large changes in the causal variables are expected over the forecast horizon, and (4) the causal variables can be accurately forecast or controlled, especially with respect to their direction. [Armstrong demo.]. Most popular among causal methods are regression based methods. *Regression analysis* is used to estimate the relationship between a dependent variable and one or more causal variables. It is typically used to estimate relationships from historical (non experimental) data. It is likely to be useful in situations in which three or fewer causal variables are important, effect sizes are important, and effect sizes can be estimated from many reliable observations that include data in which the causal variables varied independently of one another (Armstrong (2012).

Important principles for developing regression models are to use prior theory, and not statistical fit, for selecting variables. They must discard those variables that contradict with prior evidence on their influence on the forecast value. The models must be simple in terms of the number of equations, number of variables. There is evidence that regression models tend to over-fit data, so forecasts have to be "adjusted", to improve out-of-sample forecast accuracy, particularly when one has small samples and many variables.

One of the main advantages of the causal approach is that it has explanatory power. Regression based methods are appropriate when one needs to forecast what will happen using different assumptions

about the environment. Additionally, the Causal approach helps the researchers to do a "what if" forecasting, by examining the relationship between demand variables.

Regression based techniques suffer from several disadvantages, which do not make it the preferred method of choice (Armstrong 2012) in particular for sales forecasting. Even for modest requirements of forecast accuracy, these are often very difficult to implement and validate. The amount of priori information required to get a good accuracy forecast with regression is much higher than with other methods. This is because we need good estimates of all the parameters of importance and in most cases only sales figures for the past are available. These involve increased cost and time requirements in order to gather and analyse the data (Frechtling, 1996). Further we need experts with adequate knowledge of the variables and to be able to use complicated software programs (Witt and Witt, 1992).

Times Series Models (Exploratory Method)

Time series forecasting is based on the idea that the history of occurrences over time can be used to predict the future. Time series models base their prediction of the future upon past values of a variable and/or past error without attempting to discover the factors affecting the behavior of the series. A time series is a sequence of observations that are collected, observed or recorded at successive intervals of time. An intrinsic feature of a time series is that, typically a set of adjacent observations are dependent. Time series analysis is concerned with the technique of analysis of this dependence. This understanding will then help in providing a forecast for future value of time.

Unlike explanatory methods, times series forecasting do not attempt to learn factors that influence the final forecast value. These methods treat the system as a black box. The objective of times series methods is to discover a pattern in the historical data series and extrapolate that pattern to the future.

Treating the system as a black box is useful in two types of scenarios. One, where the system under consideration is extremely complex, and thus it is extremely difficult to measure the factors that influence it. The second most commonly occurring scenario is when the main concern is to only generate a forecast, with no need for any explanation of why that forecast value was generated. Sales forecasting for inventory management normally falls into this second category.

An important issue for times series methods is the availability of reliable data. The more reliable the data is, more accurate will be the final forecast. Our area of interest is to forecast sales data. The availability of extensive retail scanner data means that reliable data can be obtained for existing products. Scanner data

are detailed, accurate, timely and inexpensive. Further there is vast empirical evidence to suggest that relatively simple extrapolation methods perform as well as more complex methods. (Makridakis *et al.* 1984; Armstrong 1985).

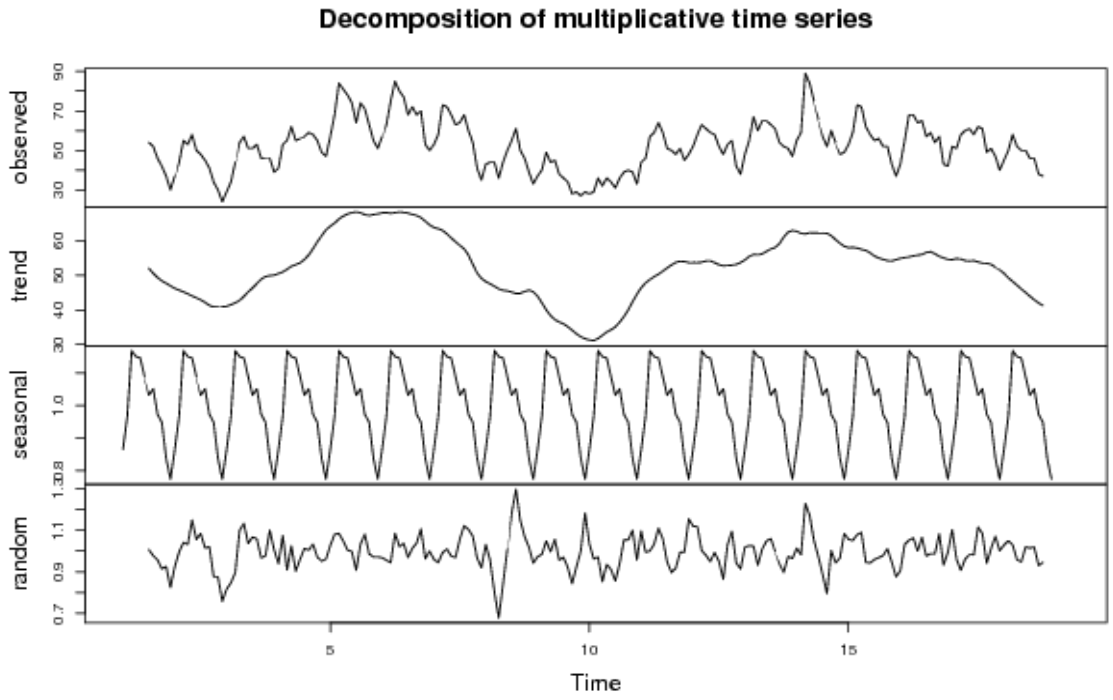
In this report the focus of research is thus restricted to the use of times series forecasting techniques.

2.2 Brief Overview of Times Series Methods

A time series is a sequence of observations taken sequentially in time. An intrinsic feature of time series it that, typically, adjacent observations are dependent. Time series analysis is concerned with the technique of analysis of this dependence. The main objective of the time series analysis is to model a process, which is generating the data, to provide compact description and to understand the generating process. Time series techniques all have the common characteristic that they are endogenous techniques. This means a time series technique looks at only the patterns of the history of actual sales (or the series of sales through time—thus, the term time series). If these patterns can be identified and projected into the future, then we have our forecast. This means that time series techniques look inside (that is, endo) the actual series of demand through time to find the underlying patterns of sales. [mentzer sage pub]

Almost all time series techniques examine one or more of only four basic time series patterns: trend, seasonality, trade cycle and noise (irregular component). ***Trend***: a long-term monotonic change of the average level of the time series, ***Trade Cycle***: a long wave in the time series, ***Seasonal Component***: fluctuations in time series that recur during specific time periods and Residual ***Irregular component*** that represents all the influences on the time series that are not explained by the other three components. Since the trend and trade cycle are changes to level most methods treat them together as a trend component. Figure 1.2 illustrates these four patterns broken out of a monthly time series of sales for Housing. As decomposition forms a vital part of our research, this topic will be further discussed in chapter 3. (Related work)

Times series forecasting models can be broadly classified into two categories. One group called “averaging methods”, in which all observations (times series values) are equally weighted. The second group applies unequal weights to past data, typically decaying in an exponential manner as one goes from recent to distant past.



2-2 Decomposition of a Series HSales

2.2.1 Averaging Models

Simple Average method uses the mean of all the past values to forecast the next value. This method can definitely be seen to be of no use in a practical scenario. This method can at best be used when the times series has attained some level of stability and no longer is dependent on any external parameters. This would happen in sales forecasting, only when the product for which a forecast is needed, is at a mature stage in its life cycle.

$$F_{t+1} = \frac{1}{t}\sum_{i=1}^{t} Y_i$$

Where F is the forecasted value, at instant of time$t = 1,\ t$ is the current time and Y_i is the value of the series at time instant$i$.

It is more reasonable to assume that the recent points in the past are better predictors than the full history. This is particularly true in the case of sales forecasting. Every product has a life cycle, initial stage, a middle volatile period and a more or less stable mature stage and an end stage. So a better method of forecasting would be to use **Moving Averages**.

The moving average approach calculates an average of a finite number of past observations and then employs that average as the forecast for the next period. The number of sample observations to be included in the calculation (order of the process) of the average is specified at the start of this process. The term **Moving** average refers to the fact that, as a new observation becomes available a new average is calculated by dropping the oldest observation in order to include the newest one. A Moving Average of order'k' $MA(k)$ is thus calculated as:

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} Y_i$$

$MA(3), MA(5), MA(12)$ are commonly used for monthly data, whereas $MA(4)$ is normally used for quarterly data. $MA(4)$ and $MA(12)$ would average out the seasonality factors in quarterly and monthly data respectively. The basic advantage of these moving average methods would be that their data requirement is very small. Since it can be adjusted to reflect the observable pattern in data to some extent is better than a simple arithmetic mean. The major disadvantage is that it assumes data to be stationary. It assumes a very simple relationship between past data. No contribution is used from early data and all past data is considered equally.

2.2.2 Exponential Smoothing Models

An extension to the moving average method is to have a weighted moving average. It is logical and reasonable to assume that more recent values of the series, will contribute more information to the forecast of the next value. Therefore while forecasting future observations; it is more appealing to put greater weight on the most recent observations. The class of Exponential Smoothing methods does this. This class of techniques consists of a huge range of methods, ranging from Simple Exponential Smoothing (SES) used for data with no trend or seasonality, to the sophisticated Holt Winter's method which is able to provide forecasts for data that exhibit both seasonality and trend. In all these methods the observations are weighted in an exponentially decreasing manner as they become older. The decaying of weights depends on a set of smoothing parameters which are to be determined explicitly.

(a) ***Simple Exponential Smoothing:*** Used for data with no trend or seasonality. The standard equation is as follows:

$$F\,(t+1) = \alpha\, D\,(t) + (1 - \alpha)\, F\,(t-1)$$

$\boldsymbol{D}\ (\boldsymbol{t})$ = Actual value at time t

$\boldsymbol{F}\ (\boldsymbol{t}+\boldsymbol{1})$ = Forecast value for time $(t+1)$

$\boldsymbol{\alpha}$ = Smoothing constant ($0 <= \alpha <= 1$ & $F(0) = D(1)$ or user input)

The two quantities are required to start off the process are the, Smoothing Constant (Alpha =α) and the starting value$D(1)$.

(b) ***Holt's Method***: This method is used when a series has no seasonality and exhibits some form of trend. the k step ahead forecast function for a given times series X is given by

$$X_{t+k} = L_t + k\,T_t$$

Where L_t is the current level and T_t is the current slope, the update equations for level and slope are

$$L_t = \alpha\, X\, t_t + (1-\alpha)\,(L_{t-1} + T_{t-1}) \quad \text{Where } 0 < \alpha < 1$$

$$T\,t_t = \beta\,(L_t - L_{t-1}) + (1-\beta)\,T_{t-1}) \quad \text{Where } 0 < \beta < 1$$

Reasonable starting values for level and slope are, $L_1 = X_1$ *and* $T_1 = X_2 - X_1$.

The values of α and β can be found by sequential search.

(c) ***Holt Winters' Method*****:**

The Holt-Winter's (HW) method is one of the best known forecasting techniques for the time series that has both trend and seasonal components. There are two variants of this method, additive and multiplicative. The seasonality is multiplicative if the magnitude of the seasonal variation increases with an increase in the mean level of the time series. It is additive if the seasonal effect does not depend on the current mean level of the time series. Since sales time series display seasonal variations the multiplicative method is described here. The Holt-Winter method uses exponential smoothing of *level* (S_t), *trend* (T_t) and *seasonal index* (I_t) for forecasting the given series,X_t . For multiplicative seasonality, the model assumes the form

$$S_t = \alpha\left(\frac{X_t}{I_{t-c}}\right) + (1-\alpha)\,(S_{t-1} + T_{t-1})$$

$$T_t = \beta\,(S_t - S_{t-1}) + (1-\beta)\,T_{t-1}$$

$$I_t = \gamma\left(\frac{X_t}{S_t}\right) + (1-\gamma)\,I_{t-c}$$

Here, c is the number of observation points in a cycle (c = 12 for monthly data). α , β and γ are the smoothing constants with values between 0 and 1. The forecast at time t for the value of the series at time $t+i$ is $(S_t + i_t) I_{t-c+i}$.

Exponential smoothing offers a simple and better alternative to the Averaging techniques, since its data requirements are very less. These models are attractive because of their simplicity and low cost and they frequently outweigh any improvement in accuracy given by other complex methods. On the flip side it does not do any better if there is some pattern in the times series. In such cases the other higher order exponential methods have to be used. These methods are very sensitive to the initial values that are chosen for slope, level etc. If there are several turning points in the series both Averaging and Smoothing methods perform very badly. The performance of these methods also depends on the values chosen for the smoothing parameters. An extension of these simple smoothing methods, adaptive exponential smoothing methods attempt to automatically learn optimal values of these smoothing constants. [] []

2.2.3 ARIMA Models (Box- Jenkins)

In real-life research and practice, patterns of the data are unclear, individual observations involve considerable error, and we still need not only to uncover the hidden patterns in the data but also generate forecasts. The Box – Jenkins(1976) Autoregressive/ Integrated/ Moving Average, ARIMA methodology is the most sophisticated of the statistical methods, which allows us to do identify complex patterns in a times series and generate an accurate forecast. It has gained enormous popularity in many areas and research practice confirms its power and flexibility (Hoff, 1983; Pankratz, 1983; Vandaele, 1983). However, because of its power and flexibility, ARIMA methodology is a complex technique which is very slow. Further, it requires a great deal of experience, and although it often produces satisfactory results, those results depend on the researcher's level of expertise (Bails & Peppers, 1982).

The ARIMA models are a general class of forecasting models that can be used for forecasting any type of times series, stationary, non-stationary, seasonal, with trend etc. They make use of all information available within the series itself to form forecasts. They heavily rely on autocorrelation patterns in the data. The objective of ARIMA models is thus, to analyse the stochastic properties of the data on their own under the philosophy that “the data will reveal its own pattern for the future’. The basis of the ARIMA model is the ARMA model, which consists of two sorts of terms: the autoregressive terms (AR) and the moving average (MA) terms.

Autoregressive (AR) Process

$$AR(1):\ y_t = a_1 y_{t-1} + \varepsilon_t$$

$$AR(p):\ y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots\ldots + a_p y_{t-p} + \varepsilon_t$$

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of ***p*** is called the order of the AR model.

Moving Average (MA) Process

$$MA(1):\ x_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

$$MA(q):\ x_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}$$

Where, β_0 is the mean of the series, ε_{t-i} are white noise, and $\beta_0 \ldots . \beta_q$ are the parameters of the model. The value of ***q*** is called the order of the MA model.

Autoregressive Moving Average (ARMA) Process

$$ARMA\ (1,1):\ y_t = a_1 y_{t-1} + \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

$$ARMA\ (p,q):\ y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}$$

The general *multiplicative seasonal ARIMA (p, d, q) X (P, D, Q)* model is expressible as

$$\Phi(B)\ \emptyset(B^c)\left(1 - B^d\right)(1 - B^c)^D\ X(t) = C_0 + \theta(B)\Theta(B^c)\varepsilon(t), \qquad t = 1, 2, \ldots$$

Here C_0 is a constant, ε(t) is a sequence of independent, zero mean and normally distributed errors, d and *D* are the orders of non-seasonal and seasonal differencing for the time series and $\Phi(B)$, $\emptyset(B_c)$, $\theta(B)$ and $\Theta(B_c)$ operators are polynomials in B (the backward shift operator) with the following general forms

$$\Phi(B) = 1 - \Phi_1(B) - \Phi_2(B^2) \ldots - \Phi p(B^p)$$

$$\emptyset(B^c) = 1 - \emptyset_1(B^c) - \emptyset_2(B^{2c}) - \ldots - \emptyset_p(B^{pc})$$

$$\theta(B) = 1 - \theta_1(B) - \theta_2(B^2) - \ldots - \theta q(B^q)$$

$$\Theta(B^c) = 1 - \Theta_1(B^c) - \Theta_2(B^{2c}) \ldots - \Theta_q(B^{qc})$$

The polynomials $\Phi(B)$ and $\theta(B)$ capture the non-seasonal behaviour and $\emptyset(B_c)$ and $\Theta(B_c)$ capture the seasonal behaviour of the series. The differencing orders *d* and *D* typically have a value *0* or *1*. For non-seasonal ARIMA, *P=D=Q=0*.

ARIMA is the most general of the time series forecasting models. It is particularly useful for sales forecasting. Both Exponential smoothing and moving averages can be configured to produce the same

results using an ARIMA forecasting model. Similarly linear Regression on past sales can be done through ARIMA forecasting model. Combination of Exponential smoothing and linear regression of past sales data can also be done through the ARIMA forecasting model. If data is non-stationary, differencing is done and ARIMA model can thus be used.

ARIMA **(p,d,q)** is the general form of this model indicating the Autoregressive (AR) of order 'p'; Moving Average (MA) of order 'q'; and 'd' indicating the order of differencing required to convert the non-stationary series to a stationary one. The ARIMA models cannot be used automatically. By studying the characteristics of the times series at hand, a careful selection of the appropriate ARIMA model suitable has to be made. The skill and understanding on the part of the forecaster can significantly enhance the effectiveness of the model. Box-Jenkins methodology gives us a set of procedures for identifying a suitable ARIMA model. This is quite a powerful tool for providing good short range forecasts. ARIMA class of models can represent a variety of patterns that can occur in times series. Once model identification is done, then ARIMA process is automatic and very easily available as a package in most forecasting or statistical software.

In spite of its many advantages the ARIMA model suffers from very severe drawbacks. Sufficient investment of time and resources are needed for model selection. A relatively large amount of data is required prior, to be able to construct a good ARIMA model. Further it needs an expert to be able to analyze the pattern of the series and then suggest a suitable model. Thus model selection process will have to be done manually and this will be cumbersome to do so for each series In addition an expert may not always be available.

ARIMA models are "Black Box" in that they do not provide any clue as to why the data follows a particular pattern. As new data become available, we may need to update the ARIMA parameters, and no simple method exists. It is also true that a given series may have different sets of patterns and so, more there may be more than one model of choice. Further different forecasting competitions have also thrown up a very surprising result that, the complexity of ARIMA does not really offer a higher degree of accuracy than many simpler methods. {M1 competition}

2.3 Evaluation of Forecasting Accuracy

We need error measures for drawing conclusions about the relative accuracy of different forecasting methods. Error measure plays an important role in calibrating and refining forecasting model/method. This calibration helps the analyst to improve forecasting method. The choice of an error measure may

vary according to the situation, number of time series available and on whether the task is to select the most accurate method or to calibrate a given model. The literature describes several error measures, each trying to capture different aspects of the loss function.

In the forecasting arena, the objective is to find the best fit for the actual series using various forecasting methods. The Root Mean Square Error (RMSE), the standard error measure in statistics, has limitations which resulted in a multitude of new error measures. The main quest was for unit-free error measures, which is one of the main limitations of RMSE. Empirical comparisons of these various error measures have been done by Armstrong [10] to compare the utility of these error measures for various forecasting methods.

A brief description of few of the popular and highly recommended error measures is given below.

1. **Mean Squared Error (MSE)** : $MSE = \frac{\sum_{j=1}^{N}(\boldsymbol{observation}\,(j) - \boldsymbol{prediction}\,(j))^{2}}{N}$

This popular measure is not unit free. A further problem with this measure is, i**t** is scale dependent, which renders it useless for comparing different forecasting methods across various series. Scale dependency also makes this measure unreliable for model calibration.

2. **Root Mean Squared Error (RMSE):** $RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\boldsymbol{observation}\,(j) - \boldsymbol{prediction}\,(j))^{2}}{N}}$

This variation on the MSE possesses similar disadvantages.

3. **Mean Absolute Percentage Error (MAPE) =** $(100/n) \times \sum_{i=1}^{n} \frac{(|X'(t) - X(t)|}{X(t)}$;

Here $X'(t)$ is the forecast of point t and $X'(t)$ is the actual value at point t.

This error measure is the most widely used unit-free error measure. A disadvantage of MAPE is it is relevant only for ratio-scaled data (i.e. data with a meaningful zero). Another disadvantage of the MAPE is that it puts a heavier penalty on forecasts that exceed the actual than on those that are less than the actual. For sales forecasting this is the error measure of choice amongst most researchers. In our work we will consistently use the MAPE as the measure to compare forecasting accuracy of different methods. A variant of the MAPE, Median Absolute Percentage Error (MdAPE) is sometimes used. This measure is almost similar to the MAPE, but in MAPE. mean is used for summarization whereas in MdAPE, median is used for summarization across the different points in a series.

2.4 Combination Forecasts

The Statistical methods mentioned in the previous section have been well established and widely used for times series analysis and forecasting from a very long time. The Box Jenkins ARIMA models have been particularly popular among most researchers and forecasters in the case of ad-hoc forecasting. The main decision issue for forecasters is to decide whether to select an appropriate modelling approach for prediction purposes or to combine different individual models to form a single combination forecast. A major challenge for researchers is to be able to formulate a procedure to select the best model for a particular series, from a set of candidate models available.

Each forecasting technique is of some use in different types of scenarios [7] [9]. Model selection suffers from the instability problem, where even a small perturbation in the input series, significantly changes the accuracy of a chosen model [AFTER]. Further, there are no objective guidelines that can be used to choose any individual model and it is completely uncertain how the selected model will affect the forecasting accuracy.

In this scenario, a combination forecast has become a de facto standard as a means to improve forecasts and to alleviate the risk of selecting unstable models. Combining has a long history that predates its use in forecasting. In 1818, Laplace claimed "In combining the results of two methods, one can obtain a result whose probability law of error will be more rapidly decreasing" (as quoted in [4]).

2.4.1 Model Selection vs. Model Combining

Model selection is a highly non-trivial task and has received considerable attention by various researchers. Different approaches have been proposed and studied. De Gooijer, Abraham, Gould, and Robinson (1985) provide a comprehensive survey and review of model selection methods. For the purpose of selecting a model, the approach of using statistical hypothesis testing techniques has several difficulties. Statistical techniques need to derive several characteristics of the series under consideration before they can choose an appropriate forecaster. Research reports three common methods for multiple candidate model selection (Yu Lean). The first is graphical inspection together with examination of simple summary statistics (such as autocorrelations (AC) and partial autocorrelations (PAC)), which is very useful for preliminary modeling analysis. The second is hypothesis testing, a formal technique for model selection in statistics. The third method is to use one of the formal model selection criteria available, such as Akaike Information Criterion , AIC [2] or Bayesian Information Criterion (BIC)[3] etc. However, all these methods suffer from some major defects. The first method is subjective and uses human inspection capabilities and thus is too informal. The difficulties with the second approach are due

to the challenging issue of multiple testing (Zou Yang). Yang in his research has also shown that a major drawback with model selection is its instability. For example, when the number of observations is small or medium, it is extremely difficult to differentiate between models that are very close to each other, since the model selection criterion values are also very similar. In this case, the choice of the model with the smallest criterion value is unstable. Even a small change in the data may lead to a different model being chosen. Thus forecasts based on the chosen model become highly variable [5, 6]. Instability of model (or procedure) selection has been addressed by several researchers and they use the term "model uncertainty" to capture the underlying difficulty of choosing the best model. (e.g., Chatfield, 1996)

To reduce variability in model selection, researchers have turned to combining or mixing the different candidate models for forecasting applications. A variety of literature on combined forecasting is reported. In a 'Combination Forecast', instead of trying to choose the single best model, an attempt is made to identify a group of models which, in conjunction, help to improve forecast accuracy. Combination forecasts offer diversification gains that make it attractive to combine individual forecasts rather than relying on forecasts from a single model.

Combination forecasting has been popularized by early forecasters such as Bates & Granger (1969) and also Granger & Ramanathan (*1984). They point out that when the objective is to have as accurate a forecast as possible, values from discarded forecasting models still contain useful information about the underlying behaviour of the series. Clemen (1989,) in his seminal paper summarizes the simulation and empirical evidence in the literature on forecast combinations, and declare "The results have been virtually unanimous: combining multiple forecasts leads to increased forecast accuracy.... in many cases one can make dramatic performance improvements by simply averaging the forecasts."

Later, Makridakis and Hibon (2000) conducted a series of competitions the M1, M2, M3-competition which involved forecasting several thousands of time series of different types. While analysing the results of the M1, M2 M3 competitions, several forecasting experts spent time and knowledge designing and tuning methods with different degrees of complexity only to come to the same conclusion: "No single best method that works well on all time series", "The accuracy of the combination of various methods outperforms, on average, the specific methods being combined and does well in comparison with other methods." [8]. Similarly, Stock and Watson (2001, 2004) undertook an extensive study across numerous economic and financial variables using linear and nonlinear forecasting models and found that, on average, pooled forecasts outperform predictions from the single best model, thus confirming Clemen's conclusion.

Typically, sales time series encounter many time varying conditions, and several change points which leads to the series exhibiting multiple Patterns at different periods of time. This will result in aggravated parameter estimation errors and model misspecification and thus it may not be possible to find, a single best method for forecasting the entire series. [6][17]. In such cases, to insure the demand forecasts against vagaries of a single model, a good strategy is to develop several decently performing forecasting models, and combine their forecasts. [3] [5] [8] [9. Research has also indicated, that individual forecasters maybe be affected differently by change points in the times series. There may be slow and fast adapting models to the change points. Since it is typically difficult to detect change points in times series without prior information, combinations of forecasts from models with different degrees of adaptability will outperform forecasts from individual models. [5]

A theoretical justification of forecast combination by considering the problem from a Bayesian model averaging perspective is provided by (Timmermann 11). When there is an inherent uncertainty in choosing the correct model that will generate the given time series, several models have to be tested, and their forecasts averaged according to the likelihood of the underlying models. Lean et al. provide another perspective to the issue of combining vs. selection. They introduce two measures "prediction performance stability" and "robustness" to characterize a set of forecasting models. If both these measures are particularly large in one forecaster as compared to all other models, a decision towards model selection as opposed to combining is advocated. In sales series research has shown that it is not possible to find one single model that has high stability and robustness for the entire series hence combination may be the alternative.

However, all the above arguments in favour of a combination forecast do not necessarily mean that model selection is useless. If we are indeed able to identify the true model that describes the series perfectly, it will always lead to better forecast accuracy. As Yang [5] claimed, in the time series context, combining does not necessarily lead to prediction improvement when model selection is stable.

In their paper (Hibon IJF) present a detailed empirical research to try and answer two important questions; first, whether the best possible combinations have better forecasting accuracy than the best possible individual forecasts, and second, when it is not possible to identify one best individual forecast or combination, is it less risky to combine than to select one. Their empirical results indicate that the best combinations are, on average across series, no better than the best individual forecasts. On the other hand when it is not known which individual forecasting method is the best, selecting among combinations leads to a significantly better performance than that of a selected individual method. Thus it can be

concluded that the advantage of a combination forecast is that selecting among combinations is less risky than selecting among individual forecasts.

Further when MAPE is the preferred choice of error measure, it can be seen that MAPE for a combined forecast will never exceed the larger MAPE value. [10]

By far the most compelling motivation comes from the fact that, enterprises would like to automate and speedup the forecasting process. Model selection will need human expert intervention and also a large amount of resources with respect to time and cost to be effective. Further the expert will expect a large amount of prior data and parametric information to be available. A combination forecast will use a number of simple models, that can be treated as a black box and what is needed is a way to combine the results for the participating methods.

2.5 Some Forecast Combination Techniques

Combination forecasting techniques pose three major issues. The choice of methods used to address these issues affects the forecasting accuracy of the final forecast. The following three issues have to be addressed:

1. Which forecasting models will participate in the combination?
2. How many of these methods will participate?
3. What is the appropriate algorithm for combining the forecasts?

Over the years different researchers have devised different answers to the above questions, with varying results. Forecast combination was pioneered in the sixties by Bates and Granger [4]. Since then many developments occurred in this field, and several review articles have appeared [3] [5] [6] [7] [8].

One of the favourable features to have in a forecast combination system is the diversity of the underlying forecasting models [7]. This will act as a buffer against being too focused on a narrow specification. Diversity can be achieved by using different forecasting models, different model specifications, or different time series pre-processing techniques. This will help to capture most of the vagaries inherent in a times series. Most researchers agree that individual forecasts in a combination should differ from each other to produce a result that improves upon individual forecasts. In general, it is desirable that the forecasts to be combined are as accurate as possible, wherein weaknesses in one forecast should preferably be compensated by the others.

More recently however, other approaches than just using different methods have been pursued in order to try a number of different individual forecasts. One method was proposed in (Granger and Yeon)] under

the name of thick modelling. The general idea here is to generate different models using the same functional approach by varying one or more parameters used in the building or forecasting process of the model. This has shown to decrease model risk and improve forecasting performance.

As an answer to the question of how to combine the models, most of the literature concentrates on five different traditional forecast combination methods. These include

- **Simple average**: The available forecasts are averaged.
- **Simple average with trimming**: The forecasts are averaged as well, but only the best 80% are taken into account.
- **Outperformance model**: Weights for a linear combination are assigned based on the number of times a forecast performed best in the past [18 lemke].
- **Variance-based model**: Weights are assigned in relation to past error variance [19 lemke].
- **Optimal model**: Weights are calculated according to [20 lemke], taking covariance information into account.

All of the methods have strengths and weaknesses as reviewed in [21 lemke]. The simple average with and without trimming has the reputation of being notoriously hard to beat. Scott and Wallis provide an explanation to this "forecast combination puzzle" by extensive analysing several combination techniques to conclude that simple combination schemes tend to perform best [13] [8]. Also, some approaches, such as shrinkage, that temper away some of the complexity, have been found to give favourable performance [5]. Trimming out bad forecasting models also tends to improve performance [5].

Another popular method for combining forecasts, Bayesian model averaging (BMA), is a standard statistical method for combining predictive distributions from different sources. The BMA forecast is a linear combination of the forecasts of the participating forecasters, where the weights are equal to posterior probabilities of the models generating the forecasts and reflect the models' relative contributions to predictive skill over the training period. The BMA weights can be used to assess the usefulness of the participating forecasts models, and this can be used as a basis for selecting them. It provides a rigorous statistical foundation where the weights assigned to the different forecasts arise naturally as posterior probabilities of the models and the combined forecast has appealing optimality properties given the set of models considered.(Min and Zellner (1993), Madigan and Raftery (1994)). Each model is evaluated, and assigned weights indicating its relevance. Every model contributes to the final forecast according to its weighting. The basic idea is that for any given forecast there is a "best" model, but we do not know what it is, and our uncertainty about the best model is quantified by BMA. While BMA is an intuitively attractive solution to the problem of accounting for model uncertainty, implementation of BMA presents

several difficulties. The number of terms in to be computed can be enormous; rendering exhaustive summation infeasible The integrals implicit in the equation are also in general be hard to compute. Specification of the prior distribution over competing models is challenging. The method is inherently statistical and very difficult to automate.

Another method often quoted in literature is the **AFTER m**ethod. The algorithm AFTER, computes a convex combination of the individual models for a better performance of prediction. The weights are sequentially updated after each additional observation. The weights for each model depend on its earlier performance. This approach differs from the earlier discussed BMA approach primarily, in the fact that no prior distributions are considered for parameters in the models. Yang (2001b) proposed this algorithm AFTER to combine different forecasts. To combine the forecasting procedures the AFTER algorithm looks at their past performances and assigns weights accordingly. After each additional observation, the weights on the candidate forecasts are updated. The algorithm is called Aggregated Forecast through Exponential Reweighting (AFTER).

Both the above mentioned methods are weight based methods. They use forecasts from all the participating models and these are used to form a weighted combination forecast.

Relatively recently, Aiolfi and Timmermann [7], [9] have also, found evidence of persistence in out of sample performance in forecasting models. This means that more or less a good or bad forecasting model stays good or bad at the later part of the series. Therefore, for a forecast combination system to be superior, the underlying forecasting models have to be good individually. Based on their empirical results that indicated that forecaster normally keeps performing in the same way instead of changing its performance, they group the forecasters into two or three clusters using a k-means algorithm based on their past performance. Forecasts are then pooled within the groups before combining them with one of the following strategies: selecting the previously best performing cluster and averaging the forecasts contained in it; excluding the cluster that performs worse and averaging forecasts from the other clusters; combining forecast averages of each of the clusters using least squares regression or doing the same as in the previous approach but shrink weights towards equal weights.

Our work is similar to the first two of the above methods. It uses a combination of “rank based” combining and data mining. Rank based combining will be described in the next chapter.

2.6 Review of Data Mining For Forecasting

In recent times there has been an increasing interest in using data mining based methods to identify the best combination of methods that can increase forecast accuracy [15], [25], [32]. Data mining models

have established themselves in the last decade as serious competitors to classical statistical models in the area of forecasting. Early candidates have been neural networks and genetic algorithms. Subsequently, research has extended to other models, such as support vector machines, decision trees, clustering, pattern mining, and others.

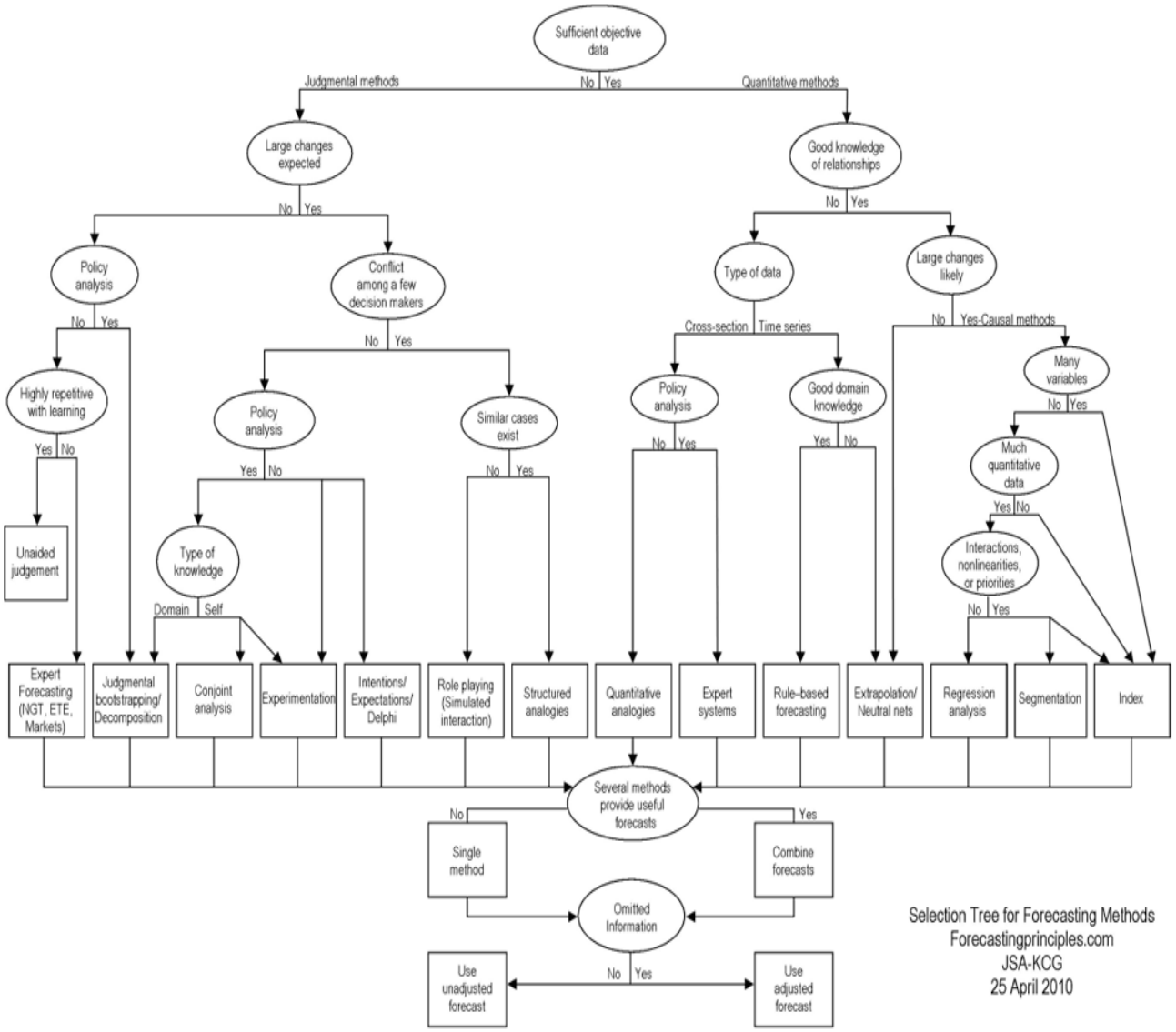
Neural Networks and Genetic Algorithms are commonly used in regression based forecasting to evaluate optimal values of parameters. [6], [19], [35]. One important issue when using ARIMA models in a combination forecast is to decide what the orders of ARIMA models need to be considered. The values chosen for p, d, q and P, D and Q dictate the goodness of the models for forecasting. To resolve this issue, (Cortez etc) used a Meta genetic algorithm, which encoded the genes with parameters of the ARIMA models. At the end of the GA procedure, optimal values for the parameters were identified. The first 90% of the series were used as training set to tune the parameters of the population at every step. The Root Mean Squared Error (RMSE) was used as the fitness function to decide which values survived and which did not. preform comprehensive empirical comparison of the performance of several novel data mining methods on the series of the M3 competition was conducted and reported in [25]. In his seminal paper Hand, [15] discusses at length how data mining could be used effectively in forecasting. The paper makes a case for how data mining models even though empirical in nature are useful, particularly when it is extremely difficult to find appropriate theoretical models for forecasting. The paper further states that Data Mining based models are a good alternative if we have a huge number of models to select from and also when the times series is very complex to pin down, as is the case for sales series.

Maharaj (1996, 1997) proposed a method of clustering stationary time series based on the p-value of a test of hypotheses that there is no difference between the generating processes of every two series under consideration. The clustering procedure has the following steps: First perform the test of hypothesis for every pair of series determining the p-value associated with the test. Use these p-values in an algorithm that incorporates the principles of hierarchical clustering but will only group together those series whose associated p-values are greater than some predetermined significance level (for example 0.05 or 0.01). This research then creates a pooled series for each cluster. Further normal ARIMA models are used for forecasting for the entire cluster. This work has been extended in [18], where clustering has been used to group items with similar demand and form a combined forecast for these series.

Till date very little work has been reported in the use of association mining (Frequent Pattern Mining - FPM) for sales forecasting. Research in this area primarily limited to intra-series mining which focuses on finding frequently appearing patterns within a single time series itself and inter-series mining which discovers the strong inter-relationship among several series [17], [26]. We propose the use of Frequent

Pattern mining to learn clusters of the best or worst models in the early part of the sales series (training phase). We further employ these models to forecast the later part of the series (testing phase).

.



We take the viewpoint that, in general, the 'true' model may or may not be in the candidate list and even if the true model happens to be included, the task of finding the true model can be very different from that of finding the best model for the purpose of prediction. We are not against the practice of model selection in general. Identifying the true model (when it makes good sense) is an important task to understand relationships between variables. In linear regression, it is observed that selection may outperform combining methods when one model is very strongly preferred, in which case there is little instability in selection. In the time series context, our observation is that, again, when model selection is stable, combining does not necessarily lead to any improvement.

3 Existing Work

The selection and implementation of a proper forecast methodology for customer demand is always an important issue for most enterprises, since the productivity of the entire enterprise can rely on the accuracy of the forecast. A significant forecast error either below or above the actual value, may result in the firm landing up with excess inventory carrying costs or else lost customers due to item shortages. When demand is fairly stable, e.g., unchanging or else growing or declining at a known constant rate, choosing an appropriate forecast method and making an accurate forecast is less difficult. If, on the other hand, the sales pattern is erratic, the complexity of the forecasting task is compounded.

Most sales series exhibit the latter characteristic. Employing multiple forecasting models captures the various irregularities that exist in a sales series. Two approaches to handle multiple forecasting models are model selection and model combination. The former involves choosing a single best model at each point, typically based on performance in the training phase. In model combination a subset of models are chosen and their average gives the forecast at that point. Combining, as an alternative to model selection, has been advocated since long . In many existing previous studies, particularly those analyzing the results of the M1, M2 M3 competitions, forecasting experts spent time and knowledge designing and tuning methods with different degrees of complexity only to come to the same conclusion: “No single best method that works well on all-time series”.

Typically, demand sales series have several inherent patterns and it may not be possible to find, the single best method for forecasting. The most compelling motivation for using a combination to forecast comes from the fact that, enterprises would like to automate and speedup the forecasting process. A combination forecast can use a vast number of simple sub-optimal models, and what is needed is an intelligent way to select and combine a subset of these models for a series.

Combining a set of simple and similar models decreases the risk of forecasting and increases accuracy. The use of series decomposition in conjunction with combining
was pioneered and further developed This work was characterized by the fact that a very large number of forecasters were used and a subset of these chosen for the final forecast.

Using this large number of forecasters as a base, data mining based methods were used to learn a set of good and bad models for a particular series. Using this information considerably increased the forecast accuracy. This work is discussed in the next section.

A time series is a sequence of observations that are collected, observed or recorded at successive intervals of time. An intrinsic feature of a time series is that, typically a set of adjacent observations are dependent. In Times Series based forecasting methods, the prediction is based on an inferred study of past general data behavior over time, i.e., the extrapolation method. An important requirement for such methods is to use reliable data. The availability of extensive retail scanner data from enterprises means that reliable data can be obtained for existing products.

Simple classical approaches to times series forecasting include the methods based on Averaging and Smoothing. Two of the best-known methods of forecasting seasonal data (such as retail sales) are the Holt-Winter method and seasonal ARIMA.

There is a need to evaluate the performance of a forecasting model in terms of the error it introduces. For demand sales, the most popular way to evaluate a forecasting model is to use the Mean Absolute Percentage Error (MAPE) value [3]. The MAPE
is defined as:

$$100 \times \sum_{i=1}^{N} (|\text{forecasted}(i) - \text{actual}(i))/\text{actual}(i)) \div N$$

The Holt Winter (HW) Multiplicative approach is used a standard approach for sales forecasting. All our results are compared with the accuracy (MAPE) of the HW method.

3.1 Rank Based Combining Methods

New classes of combination techniques have been proposed and researched .These are rank based methods. There has been sufficient empirical evidence to suggest that these new methods can increase accuracy of forecasts. These methods use basic ARIMA and smoothing models that are available in any statistical software.

These methods combined a decomposition followed by a combining step:

Each series is decomposed into its Trend, Seasonality and Irregular components using the expressions below. Let Xt, Tt, St and It denote the tth point (i.e., t^{th} month) of the respective series.

We define the Trend, , at a point t as the average sales in 12 consecutive months up to and including month = t.

$$T_t = \frac{\sum_{i=0}^{11} X_{t-i}}{12}$$

The seasonal component is

$$S_t = \frac{\frac{X_t}{T_t} + \sum_{i=1}^{n-1} S_{t-12i}}{n} \quad \text{Where } n = \left[\frac{t}{12}\right].$$

The irregular component is simply

$$I_t = \frac{X_t}{T_t \times S_t}$$

A method used in forecasting a single component is referred to as an atomic forecaster. A forecaster for the original series is a triplet made up of the atomic forecasters for each component. The set of such triplets is the Cartesian product of the sets of forecasters for the T, S and I components. Each such triplet of atomic forecasters (T, S, I) is called an "Expert". In this work, we use a total of 86 Trend models, 33 Seasonal models and 34 Irregular models.

These are mostly ARIMA and seasonal ARIMA models of different orders and various Exponential Smoothing Models. The Cartesian product of the Trend, Seasonal and Irregular models gives rise to 96,492 experts. An expert forecasts each point in the series and so we have 96,492 forecasts per point. The Appendix includes a list of atomic forecasters used in this work. There is a need to select a subset of good models from this large set of models to forecast the next point accurately.

A new way of selecting good models to combine, Rank based combining was introduced .These methods sort the different forecasters (96492) based on the MAPE value at that point. Some top K, of these forecasters are chosen to forecast the next point. This is repeated at every point. This method is called TopK. The important issue is how to choose the value K. In a variant of this method, DTopK (Dynamic Top K) all K values are evaluated and the best value of K that gives minimum MAPE value is chosen. These methods show a decrease in MAPE value when compared with the HW forecasts for the same series.

3.2 Use of Frequent Pattern Mining (FPM)

The rank based methods are cumbersome to use and very time consuming. The methods are trivial and do not effectively use the knowledge that exists in the times series and they fail to identify any pattern that could be exploited by a combination forecast. Data Mining is the process of discovering potentially valuable patterns, associations, trends, sequences and dependencies in data.

As an extension to the rank based method a frequent pattern mining based algorithm was used to learn a set of good and bad experts for a series. Association (Frequent Pattern) mining is a standard way to identify patterns from massive datasets. This method uses a portion of the times series as training, to identify expert sets that are consistently good or bad.

To forecast further points in the series, either the good set of experts are used or alternately, bad experts are filtered out and the surviving good set of experts are used. A brief overview of this work is presented in the next section.

3.3 Fine Grained Frequent Pattern Mining

An innovative method of using FPM, Fine Grained FPM, (FG-FPM) generates the set of "good" Trend T, Seasonality S and Irregular I experts for a series. In a similar fashion it was also possible to extract the "bad" T, S and I experts. These bad experts were filtered out of from the set of experts and the surviving experts used for forecasting. This approach was called the Fine Grained Filtered Experts approach (FG-FE). These approaches were able to learn the good and bad experts for a series using just 50% of the initial points. The forecast accuracy by using the surviving set of T, S and I experts in combination increased to about 16% over the Holt Winter accuracy.

• FG- FPM

The initial „n" points of the series are used as the training set. "n"
is the training size.

1. At point (i) / {1<= i <= n} take top 20000 experts based on MAPEs; Split each such good expert into its constituent S, T, & I experts.

2.At each point "i", we have 3 sets T(i), S(i), I(i) that consist the expert numbers (identities) of the atomic forecasters for the trend, seasonality and irregular components, that have appeared in the top 20,000 experts.

3. Each T, S, I expert may occur several times in the top 20,000. We fix a threshold, and disregard those experts that have a very low frequency of appearance in the top 20,000 experts. Similarly for the T and I experts also.

4. Using any standard frequent pattern mining such as Apriori or FP-tree based algorithm [] [], generate the set of experts that have consistently occurred in the top 20,000 experts list. We get 3 sets of consistent good experts.

5. We use the Cartesian product of these S, T and I experts Tg, Sg, Ig, to form the set of good experts for the given series.

6. These good experts are used to forecast the rest of the series. • FG- FE
In a similar fashion starting with the bottom 20,000 experts at each point, it is possible to use “n” points as training set and identify the set of bad atomic experts for the series , Tb, Sb, Ib

7. Filter (remove) these bad experts that appear in Sb, Tb, Ib from the sets T, S, I respectively. We use the Cartesian product of the surviving S, T and I experts, to form the forecasters for the rest of the series.

Using either of these above methods with a training size of about 50% of the points, the forecasting accuracy for forecasting the further points in the series, is increased by about 16% as compared to the HW forecasting accuracy.

4 Intelligent Forecast Engine

We now propose our improved design for an intelligent forecast engine that can be used in a retail business to forecast product demand sales. Retail enterprises have thousands of items on their inventory. It will indeed be extremely useful, if such enterprises were able to generate quick, timely and accurate forecasts for all these items. One way of generating accurate forecasts with minimum error would be to use the methods described earlier. But these methods would be prohibitive with respect to the time taken to identify a surviving set of good forecasting models for each item.

We propose an optimal solution to this problem by decreasing the number of times the FPM based algorithm must be used.

Our design works in two stages. In stage one we form clusters of all items in the inventory by grouping together all items that have similar sales patterns.
We use the standard hierarchical agglomerative clustering algorithm to cluster the items. We introduce a new distance measure, based on sales pattern similarity to perform the clustering.

For each such cluster of items, we pick one items and use its sales series to learn the best set of forecasting models based on methods in section 2. In stage 2, we use the same set of forecasting models for all items in that cluster. We illustrate the working and results of our algorithm using some real life datasets.

4.1 Stage 1: Clustering of Items

In this stage our objective is to group together, those series that follow similar sales patterns. Times series clustering has been a hotbed of research for some time now. Most of the work extends traditional methods of clustering, to time series also.

We use an agglomerative hierarchical approach to group items into clusters. A hierarchical clustering method works by grouping into clusters. A hierarchical clustering method works by grouping methods start by placing each object in its own cluster and then merge clusters into larger and larger clusters, until all objects are in a single cluster. A termination condition can then be defined to in a single cluster. A termination condition can then be defined to identify clusters of interest.

In order to decide which clusters should be combined at each iteration, a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between

pairs of observations), which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.
The usual distance measures like Euclidean or Manhattan distance will not be appropriate since we need to differentiate between two time series in terms of their sales patterns.

When it comes to finding similarities in the series, one most obvious idea would be finding correlation since correlation is the well-known measure used to find similarity in two series. Though that is true, good correlation coefficient would not assure us of similarity in increasing or decreasing patterns

Correlation does not imply causation. If two variables X and Y have a high index of correlation, it does not necessarily mean that an increase in X will cause an increase in Y. Thus we define our own measure of similarity, where we shall take into account the rise and fall of sales to match two series. Visually we can consider two series to be similar if their graphs follow fairly similar increasing and decreasing patterns.

Since it is clear that correlation will be of little use in this case, we develop an analytical method for the same. Though it is possible to find similarities in patterns by observing the graphs, the main challenge is to automate this process using a computer program and to put it in a standard way.

Thus we introduce a new measure of distance between two sales time series, where we take into account the rise and fall of sales to match two series. Visually we can consider two series to be patterns by observing the graphs, the main challenge is to patterns by observing the graphs, and the main challenge is to automate this process and defines a standard way to find the distance between two series. We call this new distance measure as *Sales Pattern Distance* (SPD) between two series.

4.2 Sales Pattern Distance (SPD)

We now define the steps required to calculate the SPD between two sales time series. Retail organization may have items whose units of sales differ very vastly in scale. For example, unit sales for television sets may be in hundreds per month as compared to audio CDs that may be sold in thousands or groceries that may be sold in hundreds of thousands per month.

Rather than looking just at the volumes of increase or decrease per month we use only the percentage difference as a measure. We only use the quantity in percentage by which the successive values in the time series vary. Since percentages are independent of the scales of original series, they are comparable.

Let us call the two series for which we want to compute the distance as X and Y. X_i and Y_i indicate the values of the series at point i.

1. Initialize SPD(X, Y) = 0;

2. Start i = 1; Start from the same point for both series X_i and Y_i

3. For next point, i+1, we compute the movement from i, ($|X_i - X_{i+1}|$, $|Y_i - Y_{i+1}|$).
   This is recorded as up, down, or flat, based on whether the sales figure went up, down, or remained the same. For very small changes say less than 5%, we assume the movement is flat.

4. We also record the percentage of change from points "i" to point "i+1" for both the series, X and Y

5. If the movement for both the series from point "i" to point "i+1" are not the same, we increment SPD(X, Y) by 1.

6. If movement is same for both the series, that is both X and Y record the same movement (up, down, flat), we check the percentage of change from the previous point. If the difference in % change between X and Y is within a permissible threshold, we consider that both the series have shown a similar pattern at these two consecutive points. If the change is beyond the permissible threshold we increment the count of SPD(X, Y) by 1.

7. Repeat steps 3 to 6, for the next set of points up to point " n" the desired training size. SPD(X, Y) is incremented at a point "i+1", if movement ($|X_i, X_{i+1}|$) not = movement ($|Y_i - Y_{i+1}|$) or if they are equal then difference in % change > threshold.

4.3 Hierarchical Agglomerative Clustering

The new distance measure proposed can be used to perform hierarchical clustering for all the items in the inventory. Let us assume that we have monthly sales figures (time series) for about "m" items in the inventory. These are denoted as I_1 , I_2 , …. I_m. Our aim is to cluster these series based on the SPD distance measure. Initially each series will be in its own cluster. The SPD as we have defined it earlier actually gives us a count of the number of points at which the two times series differ. We fix the threshold for the SPD(X, Y) at each step as some "x" % of the total number of points where the two series have been compared.

At every step, this threshold "x" will be increased by a small value (x + d). This results in some of the clusters at the previous step being merged into larger clusters. This is continued until at some stage all the series merge into one cluster. By evaluation of the clusters or using some domain knowledge we could decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion). In this work we choose to stop clustering when the threshold reaches a particular level of clustering when the threshold reaches a particular level of dissimilarity

4.4 The Algorithm

The analytical process that we propose for finding similarity of two series according to the increasing and decreasing patterns in them goes through following steps:

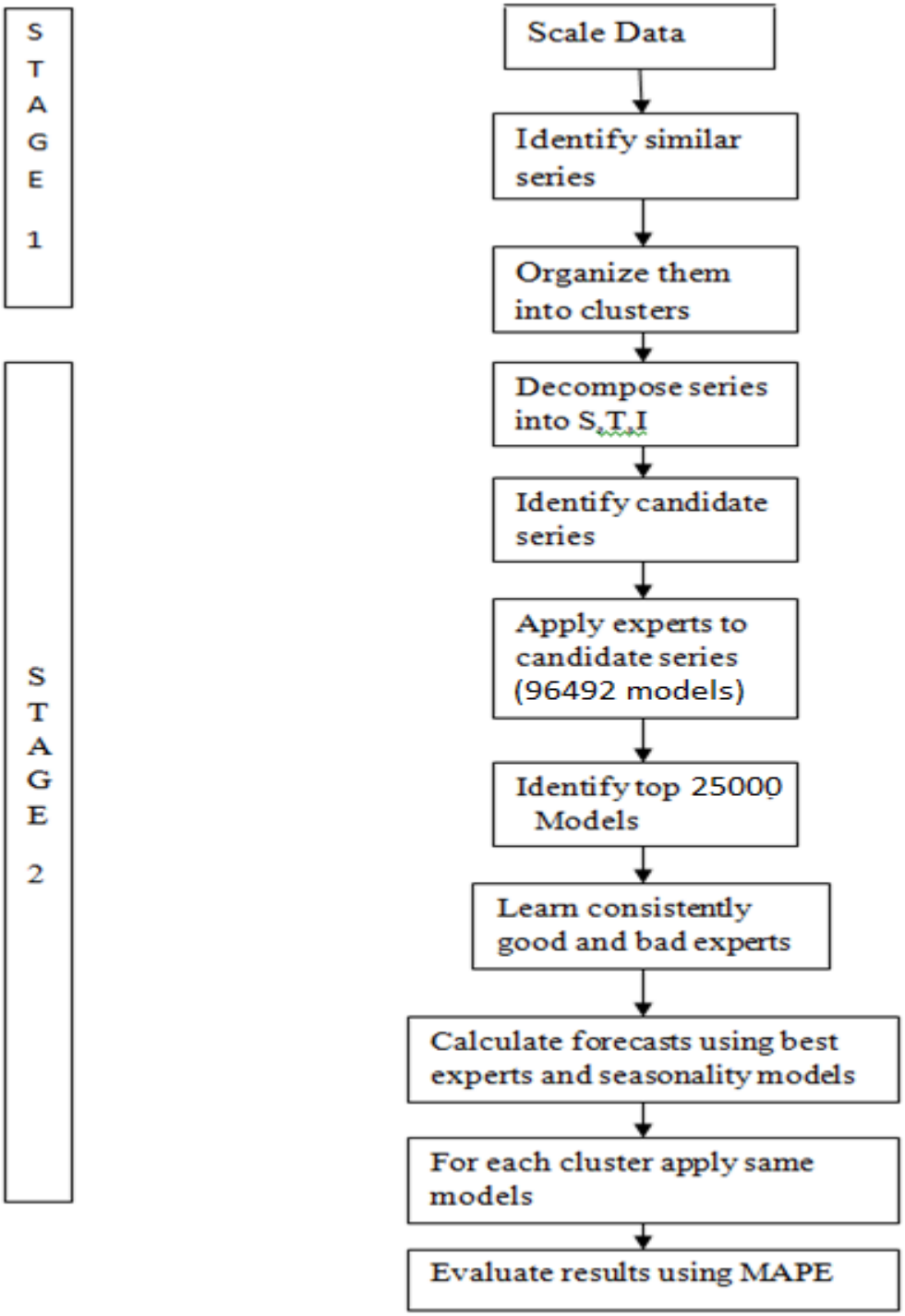


Figure 3

Step 1: *Scaling*

Retail organization may have items whose units of sales differ very vastly in scale.
For example, unit sales for television sets may be in 100s per month as compared to audio CDs that may be sold in thousands or groceries that may be sold in hundreds of thousands per month. Rather than looking just at the volumes of increase or decrease per month we will use only the percentage difference as a measure. The given series will be used to find the quantity in percentage by which the successive values in the time series vary. Since percentages are independent of the scales of original series, they would be comparable.

Step 2: *Computing Similarity between two series*.

Let us call the two series we want to compare as X & Y. X_i and Y_i indicate the values of the series at point i.

1. Use same number of points, say 'n', for X and Y corresponding to the same time period. Start from the same point for both say 'i'.

2. For next point, i+1, we can compute the movement from the previous point. This is recorded as up, down, or flat, based on whether the sales figure went up,
down, or remained the same. For very small changes say less than 5%, we shall
assume the movement is flat.

3. We also record the percentage of change. We use a threshold for percent change to decide whether the point Y_{i+1} is similar to X_{i+1} or not. If they are similar, we increment the similar count for Y by one; denote this by SimCount(X, Y).

4. Repeat steps 2, 3 up to point n.

5. If SimCount(X, Y) is greater than some predefined threshold P% of n, then
series Y is a potential candidate for similarity for series X. Add X to the
cluster that Y belongs to.

Step 3: *Putting it all together*

Let $P = \{P1, P2 \ldots Pm\}$ be the times series of the "m" items in the inventory.
Initially each item forms its own cluster. Staring at P1 we identify, all series similar to it and cluster with P1 using the method described in step 2. Next we pick any one remaining singleton

series and form clusters around it. We continue this until no more singleton series remain or the remaining singleton does not match with any other cluster.

At the end of Stage 1, we will have formed clusters of items with similar sales patterns.

4.5 Stage 2:

Finding Good Set of Forecasters for a Representative Series

We first decompose each series into its corresponding Seasonality, Trend and Irregular (S, T, I) components. For each cluster, "I", of items, we select a candidate series (CS_i).

All series considered were demand sales, for which it was found that the S component is more or less uniform. Therefore for each CS*i*, we apply 86 Trend models, 34 Irregular models and 33 models for Seasonality.
Taking the Cartesian product of each of these forecasts we should obtain about 96492 forecasts for each point.

From this pool of forecasts we identified the top 25000 (T,S, I) experts (forecasters) based on the MAPE values, for each point, in descending order. We fix a training size,"m" to decide how far into the series we need to consider to learn the consistent good or bad experts of the series.

For every point "x" in the series, starting from 25th point to mth point, we form two sets $T_{good}[x]$,$I_{good}[x]$ *and* $S_{good}[x]$. $T_{good}[x]$ consisting of the trend models that appear in the top 2500 forecasts at point x; $I_{good}[x]$ consisting of the irregular models that appear in the top 2500 forecasts at point x; $S_{good}[x]$ consisting of the seasonality models that appear in the top 2500 forecasts at point x
We use a standard frequent pattern mining algorithm, with appropriate values for support and confidence, on the sets:

$$T_{good}[x] = \{25 <= x <= m\}$$
$$I_{good}[x] = \{25 <= x <= m\}$$
$$S_{good}[x] = \{25 <= x <= m\}$$

We obtain a set of consistently good Trend experts and a set of consistently bad Irregular experts for each series.

Filtering out the bad I experts, we apply these good T and surviving good I models on our CS_i. A single forecast for T and I can be generated by taking the

mean of the forecasts of the good T and I models. The final forecast for the series is obtained by multiplying the so obtained T(mean), I(mean) and the S(mean).
For all series in the same cluster as CS_i, the forecasts for each series can be obtained in a similar fashion by using the same good T,S and I models as that of the candidate series CS_i.

To evaluate the accuracy of our algorithm, we record the MAPEs obtained for the last 24 points of each series, and compare it with the MAPE of the Holt winter forecast of the last 24 points of the series

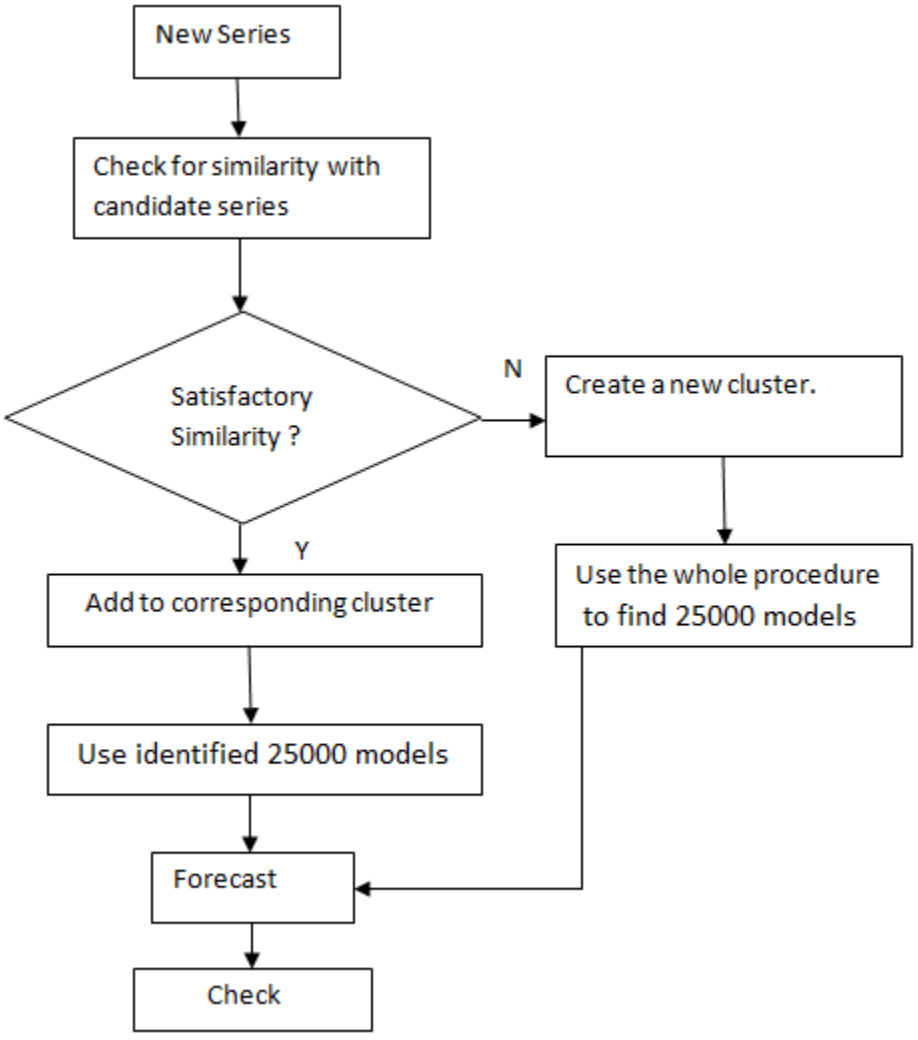


Figure 4

The above figure represents the algorithm in detail. Each series is organized into clusters, by checking similarity with a candidate series by specifying a certain threshold; or creating a new cluster if it doesn't belong anywhere.
If it finds its place in a cluster, then pre-identified models are used to forecast the time series. Else, for new clusters, the same process is applied to identify the 25000 best models, before using them for prediction.
Error measures are used to check the results.

5 Experimental Results

In this section we discuss the results of applying our algorithm on real life datasets. Empirical results indicate that for all series in the dataset, our method improves forecast results. We use the Holt Winter method as benchmark to which to compare the forecasting performance since, this method is he standard method used for demand sales.

5.1 Dataset

Our series are picked form real life dataset US retail sales from Economagic.com, available freely to subscribed users and academicians. The data was provided to us by our project guide Mrs. Vijayalakshmi M. . All series are monthly sales figures of different product from US retail data. All series have data starting from January 1980 to December 2010. We picked 25 series from this set to test the results of our hypothesis.
The Series are enumerated below:
Books:
The Books series represents the sales data over 20 years and contains about 235 points in the series. It covers a spectrum of books and publications regardless of genre or author.

Building Materials:
The Building Materials series represents the sales data over 20 years and contains about 235 points in the series. It covers a wide range of building materials, such as timber wood, bricks and cement.

- Carpets:

  The Books series represents the sales data over 20 years and contains about 235 points in the series. It covers many types of fabric floor coverings.

- Clothing:

  The Clothing series represents the sales data over 20 years and contains about 235 points in the series. It covers many types and categories of clothing, finished products and materials.

- Clothing Family:

  The Family Clothing series represents the sales data over 20 years and contains about 235 points in the series. It covers only Family clothing category in the Clothing section.

- Clothing Men:

  The Men Clothing series represents the sales data over 20 years and contains about 235 points in the series. It covers only Men clothes category in the Clothing section.

- Clothing Women:

  The Family Clothing series represents the sales data over 20 years and contains about 235 points in the series. It covers only Family clothing category in the Clothing section.

- Computer Software:

  The Computer Software series represents the sales data over 20 years and contains about 235 points in the series. It covers all computer software products over many categories such as applications, office software, games, utilities etc.

- Drugs:

  The Drugs series represents the sales data over 20 years and contains about 235 points in the series. It covers all kinds of drugs and medicines, prescription, over the counter, antivirals, antibiotics, topical and internal drugs etc.

- Electronics:

  The Electronics series represents the sales data over 20 years and contains about 235 points in the series. It covers all kinds of electronic goods, consumer durables such as PC sets, music systems etc.

- Food and Beverage:

  The Food and Beverage series represents the sales data over 20 years and contains about 235 points in the series. It covers all types of food and beverage products such as perishable items, canned food, and ready to eat.

- Furniture:

  The Furniture series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds and categories of furniture, branded and unbranded, upholstery, etc.

- Gifts:

The Gifts series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of gift items and products, soft toys, momenta, cards etc.

- Hardware:

  The Hardware series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of hardware and has products such as wiring, lighting, tools and accessories.

- Health:

  The Health series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of health and personal care products, supplements, ointments and lotions, shaving kits etc.

- Jewelry:

  The Jewelry series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds branded, luxury and affordable jewelry products, and various lines.

- Liquor:

  The Liquor series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of alcohol according to brand and alcohol content

- Motor Vehicles:

  The Motor Vehicle series represents the sales data over 20 years and contains about 235 points in the series. It covers a wide range of vehicles, two wheelers, cars, SUVs as well as second hand vehicles.

- New Car

  The New Car series represents the sales data over 20 years and contains about 235 points in the series. It covers a section of the vehicles data, specifically brand new cars.

- Used Car:

  The New Car series represents the sales data over 20 years and contains about 235 points in the series. It covers a section of the vehicles data, specifically second or third hand vehicles.

- Petrol:

  The Petrol series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of gas, such as premium quality, unleaded and branded.

- Shoes:

  The Shoes series represents the sales data over 20 years and contains about 235 points in the series. It covers a wide variety of footwear for men, women and children.

- Sports Good:

  The Sports Goods series represents the sales data over 20 years and contains about 235 points in the series. It covers a range of sports products like golf kits, racquets and baseball gear etc.

- Grocery:

  The Shoes series represents the sales data over 20 years and contains about 235 points in the series. It covers many kinds of grocery products, vegetables and fruits, and dairy.

- Paint Wall Paper:

  The Wall Paper series represents the sales data over 20 years and contains about 235 points in the series. It covers a range of wallpapers, according to designs and sizes.

5.2 Decomposition

Time Series forecasting methods often decompose the original time series into Trend (denoted as T), Seasonal (denoted as S) and Irregular (denoted as I) components. This decomposition is useful since relevant models and experts can be used for respective components, rather than using more general models.

The functional relationship between these components may have different forms. However, two simple possibilities are that they behave in an *additive* or a *multiplicative* fashion:

Additive model:

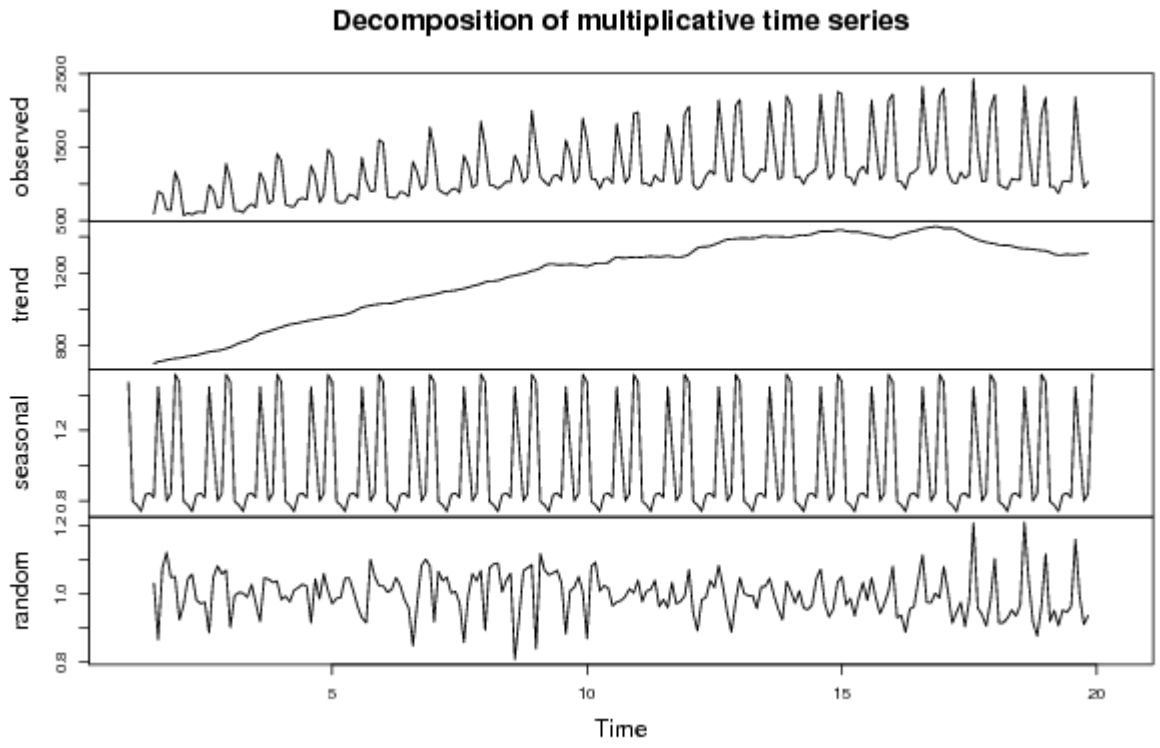
$X=S+T+I$

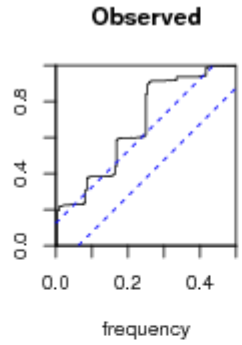
Multiplicative model:

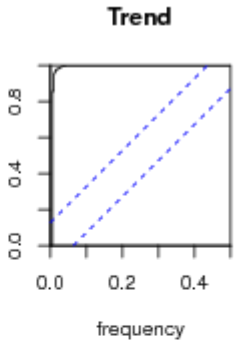
$X=S * T * I$

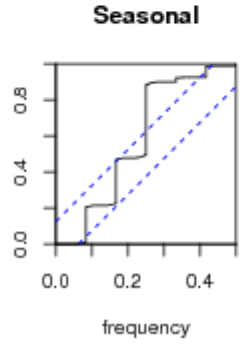
A sample of the decomposition of series is shown:
The series under consideration is the book series. The graphs below show the Trend, Seasonality and random or irregular component respectively.

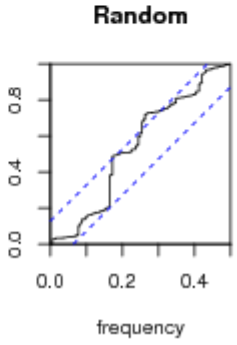
5.2.1 Books Series- Sample Results of Decomposition

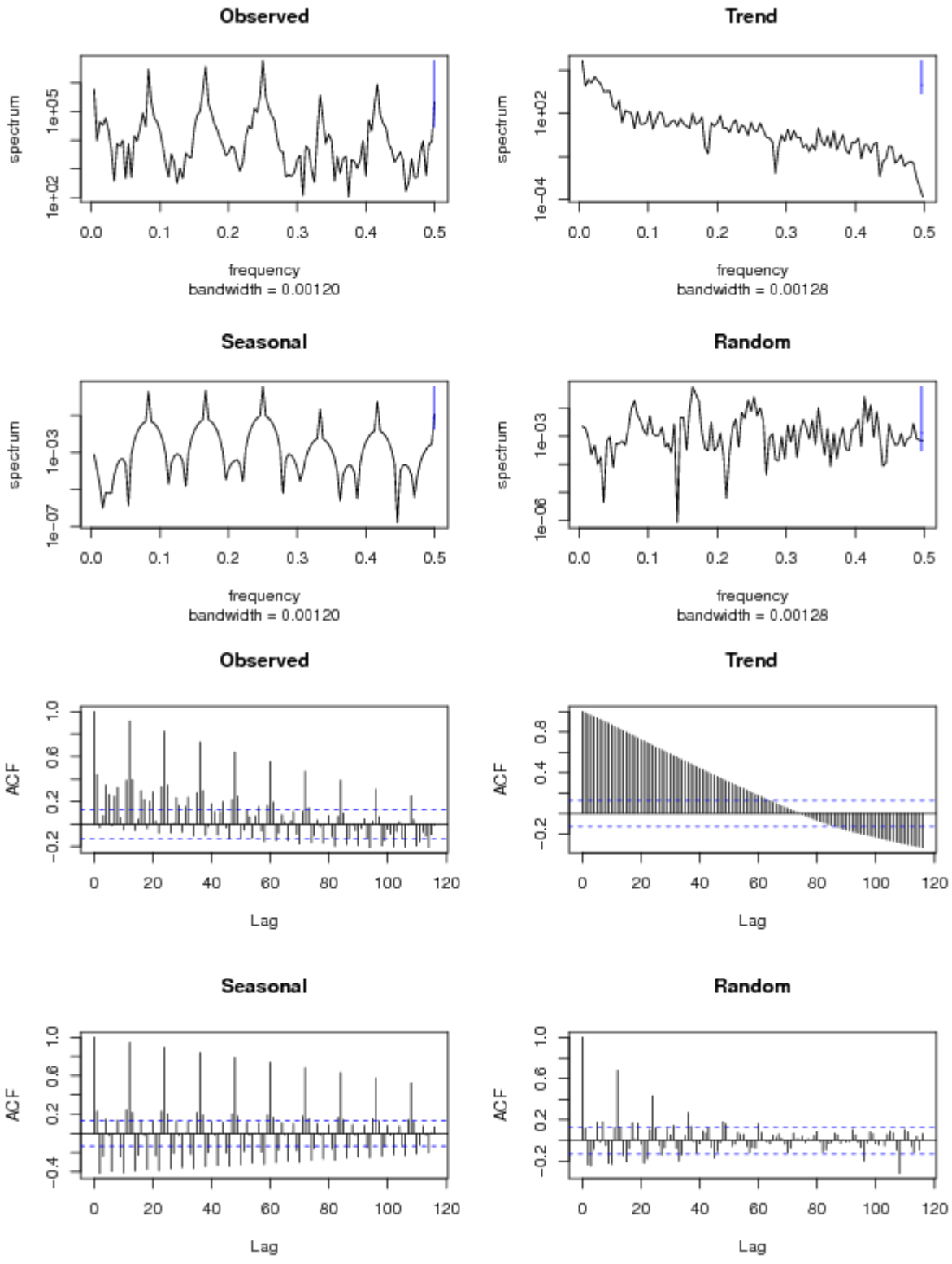











Observed
spectrum
1e+05
1e+02
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00120
Trend
spectrum
1e+02
1e−04
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00128
Seasonal
spectrum
1e−03
1e−07
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00120
Random
spectrum
1e−03
1e−06
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00128
Observed
ACF
1.0
0.6
0.2
−0.2
0
20
40
60
80
100
120
Lag
Trend
ACF
0.8
0.4
−0.2
0
20
40
60
80
100
120
Lag
Seasonal
ACF
1.0
0.6
0.2
−0.4
0
20
40
60
80
100
120
Lag
Random
ACF
1.0
0.6
0.2
−0.2
0
20
40
60
80
100
120
Lag

5.2.2 Snapshots of Excel Data

	A	B	C	D	E	F
1	t	Observati	Fit	Trend	Seasonal	Random
2	1	790	NA	NA	1.47439	NA
3	2	790	NA	NA	0.795567	NA
4	3	790	NA	NA	0.776	NA
5	4	790	NA	NA	0.739335	NA
6	5	553	NA	NA	0.83349	NA
7	6	589	NA	NA	0.843502	NA
8	7	593	627.4471	767.7083	0.817299	0.9451
9	8	895	1111.02	767.1667	1.448212	0.805566
10	9	863	838.5867	750.0833	1.117991	1.029112
11	10	647	585.7937	733.625	0.798492	1.104484
12	11	642	612.2483	727.5	0.841578	1.048594
13	12	1166	1107.218	731.25	1.514145	1.05309
14	13	999	1080.912	733.125	1.47439	0.924219
15	14	568	586.7305	737.5	0.795567	0.968076
16	15	602	576.5677	743	0.776	1.04411
17	16	583	551.2978	745.6667	0.739335	1.057505
18	17	613	624.0405	748.7083	0.83349	0.982308
19	18	619	637.1603	755.375	0.843502	0.971498
20	19	608	622.986	762.25	0.817299	0.975945
21	20	985	1111.382	767.4167	1.448212	0.886284
22	21	905	862.6701	771.625	1.117991	1.049068
23	22	669	618.1326	774.125	0.798492	1.082292
24	23	693	654.9233	778.2083	0.841578	1.058139
25	24	1275	1189.55	785.625	1.514145	1.071834
26	25	1055	1169.253	793.0417	1.47439	0.902286
27	26	636	638.9396	803.125	0.795567	0.995399
28	27	635	632.569	815.1667	0.776	1.003843
29	28	610	608.2882	822.75	0.739335	1.002814
30	29	684	690.7197	828.7083	0.83349	0.990271
31	30	726	706.5383	837.625	0.843502	1.027545
32	31	679	697.837	853.8333	0.817299	0.973007
33	32	1156	1256.927	867.9167	1.448212	0.919703
34	33	1023	977.0779	873.9583	1.117991	1.046999
35	34	733	702.5066	879.7917	0.798492	1.043407