Table of Content:

1 Introduction: 8
1.1 Problem Statement: 8
1.2 Objectives of the Study: 9
1.3 Assumptions of the Study: 10
1.4 Delimitation of the Study: 10
2 Literature Review: 12
2.1 Role of forecasting: 12
2.2 Forecasting methods: 16
2.2.1 Time Series Analysis: 21
2.2.2 Association or causal statistical techniques: 24
2.3 Usage of different forecasting models: 36
2.3.1 Forecasting accuracy and bias: 39
2.4 Forecasting method selection: 43
2.4.1 Characteristics of the data: 47
2.4.2 Type of data: 48
2.4.3 Range and frequency of forecast: 48
2.4.4 Other factors to be considered: 50
3 Methodology: 52
3.1 Gathering of the data: 53
3.2 Exploratory analysis: 53
3.3 Selection of quantitative models: 54
3.4 Evaluation of the models: 55
3.5 Combination of forecasts: 55
4 Study Results 56
4.1 Data variables and Exploratory Analysis of data used in the study: 56
4.1.1 Internal indicators 58
4.1.2 External variables: 60
4.2 Quantitative models applied 63

4.2.1 Time Series Decomposition for Commuter Rail data: 63
4.2.2 Exponential Smoothing Additive Model: 67
4.2.3 Winter's multiplicative model: 68
4.2.4 Exponential Smoothing Multiplicative Model: 70
4.2.5 Comparison of the time series models: 71
4.2.6 Box-Jenkins or ARIMA Model: 71
4.2.7 Regression and Multi-Regression model: 83
4.3 Comparison of various quantitative models applied. 99
4.4 Combination of forecasts: 101
4.5 Forecast of Ticket Sales for Commuter Rail for October 2008 to March 2010. 103
5 Conclusions and recommendations 108
5.1 Conclusions from the study conducted 108
5.2 Recommendations 111
5.2.1 Recommendations for further study: 111
6 Acknowledgements: 113
7 References 114
8 Appendices: 117
8.1 ARIMA(1,1,0)(1,0,0) model forecasts with confidence intervals ...117
8.2 ARIMA (1,1,0)(1,0,2) model forecasts with confidence intervals ..118

List of Graphs:

Graph 1: SARCC : Ticket Issues 56

Graph 2: SARCC: Trains on time 58

Graph 3: SARCC: Trains Cancelled 58

Graph 4: SARCC: Serious Crime Incidents 59

Graph 5: SARCC: Ticket Price Index 59

Graph 6: Inflation 60

Graph 7: Average Fuel Prices: Petrol and Diesel 60

Graph 8: Prices: Illuminating Paraffin 61

Graph 9: Box and Whisker Plot: Ticket Issues, Trains on time, Train Cancellations and Serious Crime Incidents 62

Graph 10: Box and Whisker Plot: Price Index, CPI and CPIX 62

Graph 11: Box and Whisker Plot: Average Fuel prices: Petrol, diesel and illuminating paraffin 62

Graph 12: Exponential Smoothing: Ticket Issues 64

Graph 13: Long Term Linear Trend 64

Graph 14: Seasonal Trend 65

Graph 15: Seasonally adjusted and detrended data 66

Graph 16: Cyclical Trend 67

Graph 17: Exponential Smoothing Additive Model 68

Graph 18: Winter's Multiplicative Model 69

Graph 19: Exponential Smoothing Multiplicative Model 70

Graph 20: Ticket Issues prior to differentiation 71

Graph 21: Differenced Ticket Issues - Lag 1 72

Graph 22: Autocorrelation: Ticket Issues 73

Graph 23: Partial Autocorrelation: Ticket Issues 74

Graph 24: Residuals of ARIMA (1,1,0)(1,0,0) 75

Graph 25: Distribution of ARIMA (1,1,0)(1,0,0) residuals. 75

Graph 26: Autocorrelation: ARIMA (1,1,0)(1,0,0) Residuals 76

Graph 27: Partial Autocorrelation: ARIMA(1,1,0)(1,0,0) residuals 77

Graph 28: Forecasts of ARIMA(1,1,0)(1,0,0)..78
Graph 29: Distribution of the residuals: ARIMA(1,1,0)(1,0,2) model 80
Graph 30: Autocorrelation residuals : ARIMA (1,1,0)(1,0,2)......................81
Graph 31: Partial Autocorrelation of residuals: ARIMA(1,1,0)(1,0,2) 82
Graph 32: Forecast of ARIMA(1,1,0)(1,0,2) 83
Graph 33: Scatter plots of Ticket Issues and Internal predictors................84
Graph 34: Scatter plots of Ticket Issues and External predictors 85
Graph 35: Comparison of forecast results for time series forecasts 100
Graph 36: Comparison of forecast results of ARIMA forecasts 100
Graph 37: Comparison of combination of forecasts of ARIMA models 102
Graph 38: Forecast of ticket sales for October 2008 to March 2010.......... 105
Graph 39: Forecast of Ticket Sales using seasonal index adjustment........ 106

List of Tables:

Table 1: Descriptive statistics: Ticket issues 57
Table 2: Means and Standard Deviations for Regression variables 63
Table 3: Seasonal trend 65
Table 4: Summary of error values: Exponential Smoothing Additive model .68
Table 5: Winter's Multiplicative Model: Error values 69
Table 6: Exponential Smoothing Multiplicative Model: Error values 70
Table 7: Parameters for ARIMA (1,1,0)(1,0,0) model 74
Table 8: Autocorrelation for residuals: ARIMA(1,1,0)(1,0,0)......................77
Table 9: Partial Autocorrelation Function Residuals: ARIMA(1,1,0)(1,0,0) ..78
Table 10: Results for ARIMA (1,1,0)(1,0,1)..79
Table 11: Autocorrelations of Residuals: ARIMA(1,1,0)(1,0,2)...................80
Table 12: Partial Autocorrelation of residuals: ARIMA(1,1,0)(1,0,2) 82
Table 13: Correlation results for ticket issues and trains-on-time: F-test 86
Table 14: Correlation results for ticket issues and trains on time: t-statistic 86
Table 15: Correlation result for ticket issues and trains cancelled: F-test 87
Table 16: Correlation result for ticket issues and trains cancelled: t-statistic 87
Table 17: Correlation results for ticket issues and serious crime incidents: F-test..88
Table 18: Correlation results for ticket issues and serious crime incidents: t-statistic 88
Table 19: Correlation results for ticket issues and price index: F-test 89
Table 20: Correlation results for ticket issues and price index: t-statistic 89
Table 21: Correlation results for ticket issues and CPI: F-test....................89
Table 22: Correlation results for ticket issues and CPI: t-statistic...............89
Table 23: Correlation results for ticket issues and CPIX: F-test..................90
Table 24: Correlation results for ticket issues and CPIX: t-statistic............90
Table 25: Correlation results for ticket issues and price of petrol: F-test.....91
Table 26: Correlation results for ticket issues and price of petrol: t-statistic 91

Table 27: results for ticket issues and price of diesel: F-test 92
Table 28: Correlation results for ticket issues and price of diesel: t-statistic 92
Table 29: Correlation results for ticket issues and price of illuminating paraffin: F-test 92
Table 30: Correlation results for ticket issues and price of illuminating paraffin: t-statistic 93
Table 31: Regression between CPI and CPIX 94
Table 32: Regression between CPIX and Fuel Price for Petrol 94
Table 33: Regression between CPIX and Fuel Price for Diesel 95
Table 34: Regression between CPIX and Fuel Price for Illuminating Paraffin 95
Table 35: Regression between Fuel price for petrol and Fuel price for diesel 96
Table 36: Regression between Fuel price for petrol and Price of Illuminating Paraffin 96
Table 37: Results of KB test for trains on time 97
Table 38: Results of KB test for CPIX 98
Table 39: Results of KB test for illuminating paraffin 98
Table 40: Durbin-Watson statistic for multi-regression variables...............99
Table 41: Comparison of Error values of different quantitative forecast models 101
Table 43: Parameters for ARIMA (1,1,0)(1,0,2) model for forecast 104
Table 44: Application of ARIMA(1,1,0)(1,0,2) model with seasonal index..106
Table 45: Comparison of forecasting errors of quantitative models 110

List of Figures:

Figure 1: Forecasting as part of business planning 13
Figure 2: Information flows of sales and business planning 14
Figure 3: Classification of forecasting techniques 17

1 Introduction:

Commuter rail transport in South Africa is provided in major metropolitan areas of South Africa and transports on average 2.2 million commuters per weekday. The service is provided by Metrorail, a division of the South African Rail Commuter Corporation Limited (SARCC) and is one of the three public modes of transport, the other being taxis and buses. Forecasting of ticket issues and therefore revenue accurately in such a consumer environment is dependent on external and internal factors influencing commuter buying patterns and travelling patterns.

1.1 Problem Statement:

Fare revenue is the end result of usage (ticket sales) of the system by commuters that use the commuter rail system on a daily basis to mainly travel between home and work. The forecasting of fare revenue the commuter rail environment for Metrorail forms an important basis for the measurement of the achievement of strategy in the organisation as well as for performance management of the regions where the service is rendered. Current forecasting of fare revenue is based on naive and judgmental forecasting that is more subjective than quantitative techniques and a more scientific base is required.

Forecasting of the fare revenue has been based qualitatively on high level trend analysis of ticket sales or issues and the judgement of the impact of external trends in the economy that impacts commuter disposal income as well as assessing the impact of internal service issues such as punctuality of trains and personal safety on a judgmental basis. The current methodology applied relies significantly on the knowledge of experienced individuals within the organisation and by developing a quantitative forecasting model more robust forecasting can be utilised. In addition there is the need to understand the factors impacting the revenue so that there is a better understanding of future inflection points.

The objective of this research proposal is to obtain the most suitable quantitative forecasting method(s) that could provide a more scientific basis for forecasting of ticket sales or ticket issues which can easily be implemented. In addition the research proposal will focus on external factors that impacts disposable income of commuters and the impact of internal service delivery issues on commuter satisfaction that can be used in conjunction with judgment of senior staff members in the organisation for forecasting. The analysis of the factors will enable the explanation of the impact of the factors on the performance of fare revenue.

1.2 Objectives of the Study:

In line with the problem statement above the objectives of the study is to:

- Determine the influence of various external factors such as economic growth or economic indicators as well as internal factors such as service delivery factors on the performance of ticket issues and
- To determine a quantitative statistical model or models that will provide the best forecasting results for ticket issues in the commuter rail environment that can be used relatively easily in the future.

The objectives will be based on the assumptions indicated in the following paragraph.

1.3 Assumptions of the Study:

Assumptions clarify the context of the study and to this end the report is based on the following assumptions:

- The assumption of continuity that some patterns displayed in the past in the sales of commuter rail tickets will continue into the future;
- It is assumed that the internal data for service delivery is sufficient indicators for commuter choices in terms of the service delivered.
- The overall results are assumed to be an indicator of regional performance;
- It is assumed that travelling patterns of commuters remain fairly constant.
- It is also assumed that a quantitative method will be able to provide explanations of trends and be of greater accuracy than the naïve methods currently employed.

These assumptions therefore limit the application of the study for future use and the delimitation of the study will be discussed in the next paragraph.

1.4 Delimitation of the Study:

The assumptions of the study discussed limit the application of the study.

This study is limited from an internal perspective to data readily available for the stakeholders of the business. The forecasting is only applicable to the Metrorail service of the SARCC and will only consider known measurable factors in the explanatory analysis of the commuter environment. As human behaviour cannot be exactly predicted the study is limited to those factors that can contribute to commuter behaviour and that can be collected in metric terms.

The study will forecast ticket issues as indicator of the sales volumes to determine fare revenue.

The factors considered by the study for prediction of the sales of tickets in future only provides the relative importance of each factor to the overall prediction and do not represent the causes of fare revenue changes. As only factors for which data is readily available is considered only the relative importance of these variables are considered. This therefore indicates that as other factors that can influence fare revenue becomes known and data can be collected for these factors, a different model will have to be constructed to take into consideration the impact of these factors on commuter decisions.

The results will further be limited to the models available in the software package Statistica version 8 of STATSOFT.

A literature review to determine the best way to approach the forecasting issue identified is discussed in Chapter 2.

2 Literature Review:

The following areas will be discussed under the literature review – Role in business planning, use of forecasting methods, various forecasting methods and factors influencing the decisions on forecasting methods.

2.1 Role of forecasting:

(Gaither and Frazier, 2002: 63) state that "forecasting is an integral part of business planning". Inputs processed through forecasting models or methods are used to develop demand estimates. These demand estimates in itself is not the sales forecast but becomes input to business strategy and resource forecasts. (Makridakis, Wheelwright and Hyndman, 1998) also see forecasting as an integral part of the decision making activities of management. (Saffo, 2007) feels that "At the end of the day, Forecasting is nothing more (or less) than the systematic and discipline application of common sense".

Figure 1, from (Gaither and Frazier, 2002: 65) illustrates the integral part of forecasting in business planning:

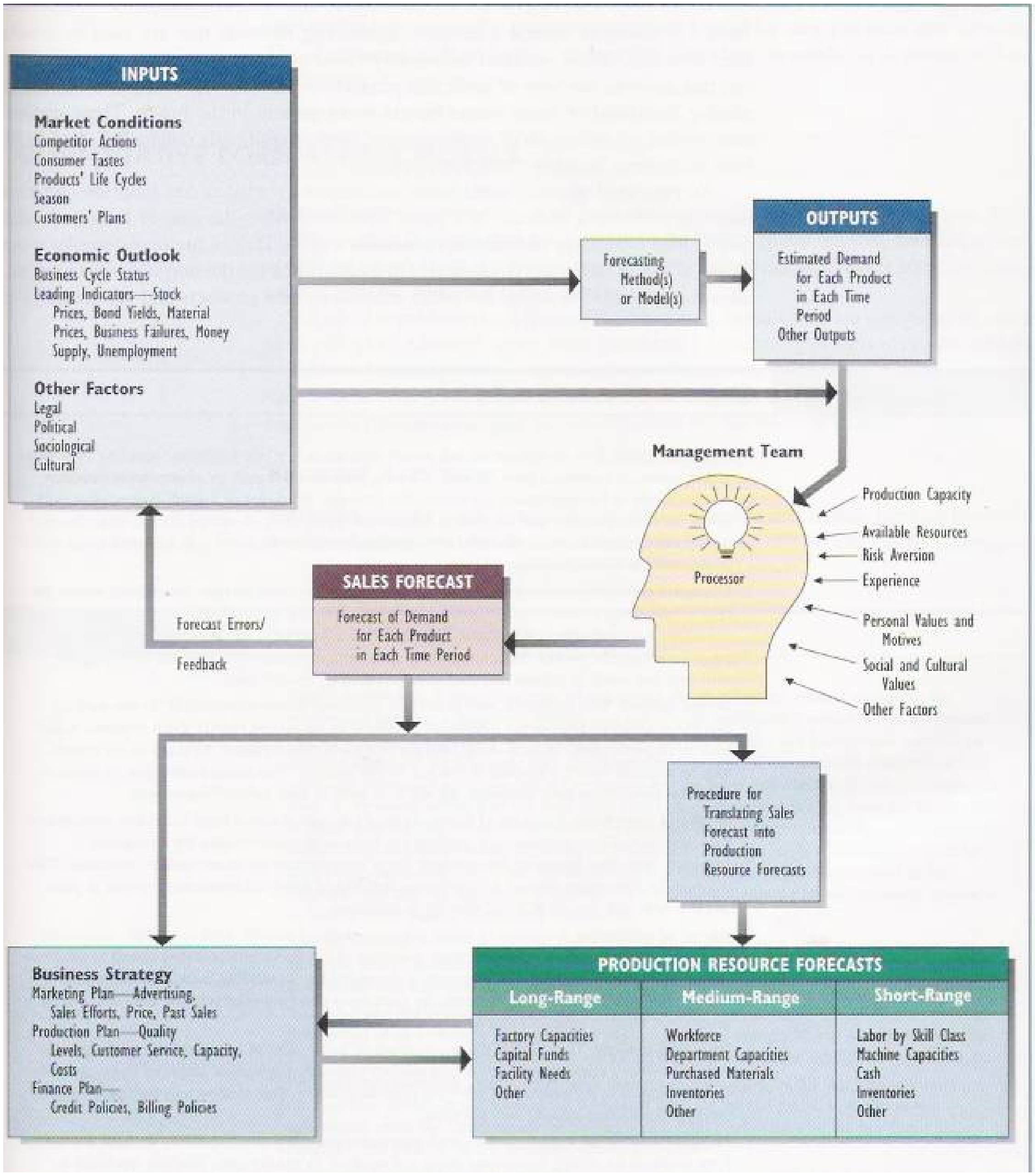

INPUTS
Market Conditions
Competitor Actions
Consumer Tastes
Products' Life Cycles
Season
Customers' Plans
Economic Outlook
Business Cycle Status
Leading Indicators—Stock Prices, Bond Yields, Material Prices, Business Failures, Money Supply, Unemployment
Other Factors
Legal
Political
Sociological
Cultural
Forecasting Method(s) or Model(s)
OUTPUTS
Estimated Demand for Each Product in Each Time Period
Other Outputs
Management Team
Processor
Production Capacity
Available Resources
Risk Aversion
Experience
Personal Values and Motives
Social and Cultural Values
Other Factors
SALES FORECAST
Forecast of Demand for Each Product in Each Time Period
Forecast Errors/
Feedback
Procedure for Translating Sales Forecast into Production Resource Forecasts
Business Strategy
Marketing Plan—Advertising, Sales Efforts, Price, Past Sales
Production Plan—Quality Levels, Customer Service, Capacity, Costs
Finance Plan—Credit Policies, Billing Policies
PRODUCTION RESOURCE FORECASTS
Long-Range
Medium-Range
Short-Range
Factory Capacities
Capital Funds
Facility Needs
Other
Workforce
Department Capacities
Purchased Materials
Inventories
Other
Labor by Skill Class
Machine Capacities
Cash
Inventories
Other

Figure 1: Forecasting as part of business planning

(Makridakis *et al.*, 1998: 4) illustrate as per Figure 2, the information flows in sales forecasting and business planning:

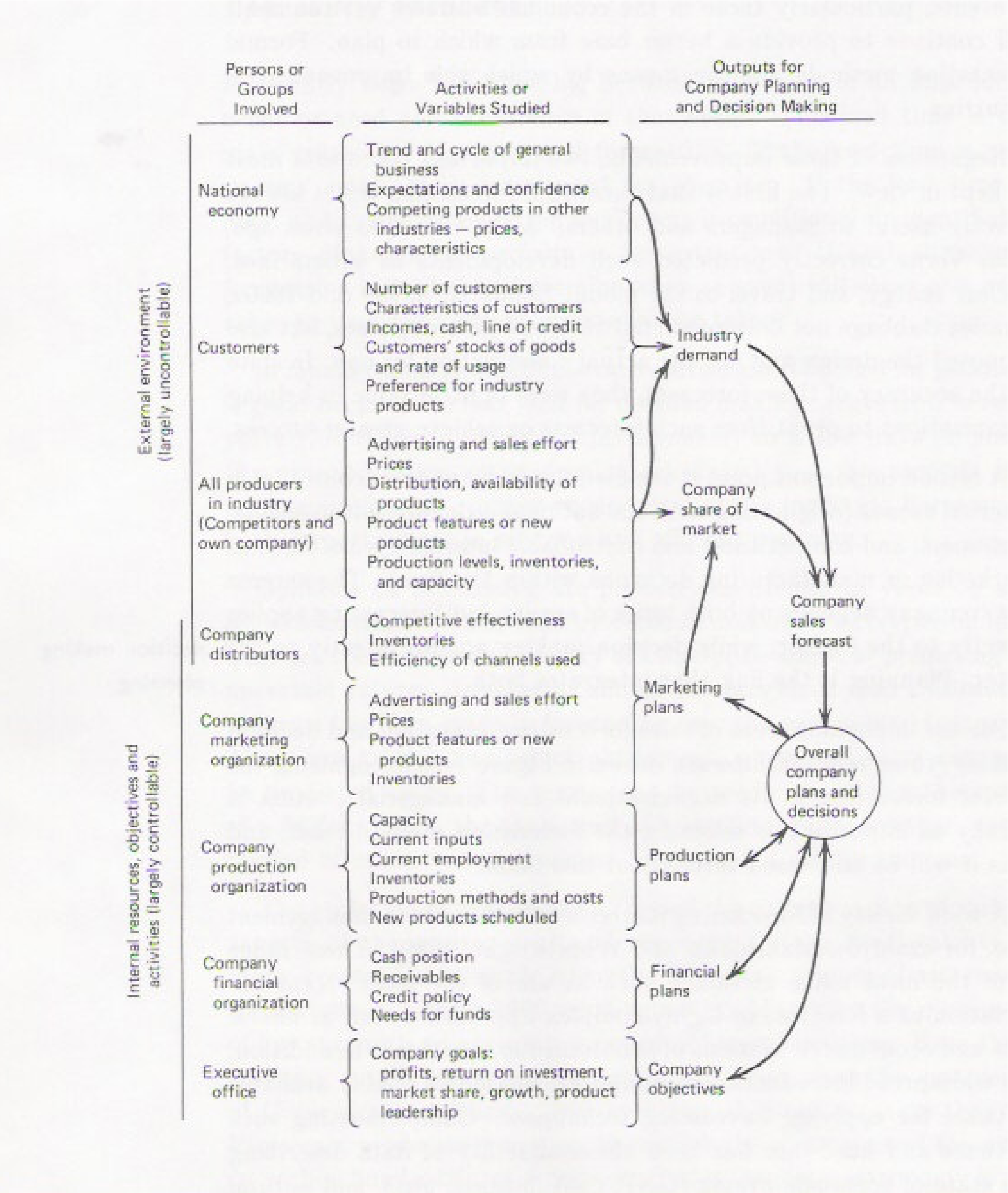

Persons or Groups Involved
Activities or Variables Studied
Outputs for Company Planning and Decision Making
External environment (largely uncontrollable)
Internal resources, objectives and activities (largely controllable)
National economy
Trend and cycle of general business
Expectations and confidence
Competing products in other industries — prices, characteristics
Customers
Number of customers
Characteristics of customers
Incomes, cash, line of credit
Customers' stocks of goods and rate of usage
Preference for industry products
All producers in industry (Competitors and own company)
Advertising and sales effort
Prices
Distribution, availability of products
Product features or new products
Production levels, inventories, and capacity
Company distributors
Competitive effectiveness
Inventories
Efficiency of channels used
Company marketing organization
Advertising and sales effort
Prices
Product features or new products
Inventories
Company production organization
Capacity
Current inputs
Current employment
Inventories
Production methods and costs
New products scheduled
Company financial organization
Cash position
Receivables
Credit policy
Needs for funds
Executive office
Company goals: profits, return on investment, market share, growth, product leadership
Industry demand
Company share of market
Company sales forecast
Marketing plans
Overall company plans and decisions
Production plans
Financial plans
Company objectives

Figure 2: Information flows of sales and business planning

As management attempts to decrease dependence on chance and becomes more scientific in the dealing of the business environment, the need for forecasting increases according to (Makridakis *et al*., 1998). In addition they also indicate that due to the interrelatedness of the various business areas a good or bad forecast will affect the entire organisation. The range of needs of modern companies requires companies to build multiple approaches to "predict uncertain events and build up a system for forecasting" (Makridakis *et al*., 1998). They also indicate that this requires knowledge and skills in the organisation in four areas namely:

- Identification and definition of the organisation's forecasting problems;
- Application of different forecasting methods;
- Procedures to select the most appropriate forecasting method for a given situation and
- The need for organisational support for applying and utilising formalised forecasting methods.

Intuit, a firm specialising in financial software utilises the principles of these two models in their process for forecasting and planning by using six steps of
1) planning and direction linked to strategic objectives and evaluation of performance,
2) volume forecasting through involvement of various teams in the company,
3) staff planning or "production planning",
4) model evaluation through simulations and revisiting of volumes and staff if necessary,
5) execution of plans or "production" and
6) post mortem that involves auditing results of the performance and improvement of the forecasting and planning model. (Ramanujan and Fisher (2006)).

(Saffo, 2007) states that forecasting looks at "how hidden currents in the present signal possible changes in direction for companies". The primary goal of forecasting according to (Saffo, 2007) is to "identify the full range of possibilities" and he sees the forecaster's task to "map uncertainty". He also indicates that users of forecasts must understand the forecast sufficiently to independently assess the quality of the forecast in order to fully account the risks and opportunities presented in the forecast. Effective forecasting provides management with understanding through exposure of unexamined assumptions in projected outcomes and revealing overlooked possibilities, as well as narrowing the decision space.

(Jain, 2008) indicates that the forecasting process includes how the data is collected and used, the analysis and treatment of data and model selection, the overlay of judgments by various functions overlay the forecasts and how these forecasts are monitored and updated.

(Makridakis *et al*., 1998) also states that no statistical method or any other approach including judgmental forecasts allow one to accurately forecast the extent of future uncertainty if the past pattern does not repeat itself. Forecasts are however needed for planning, scheduling and strategic decisions and one has to find the most rational and economic way to obtain these. The requirement for forecasts is that the forecasts must be as accurate as possible with forecasting errors as small as possible and estimated as realistically as possible.

2.2 Forecasting methods:

(Pycraft, Singh, Phihlela, Slack, Chambers, Harland, Harrison and Johnston, 2003) classifies forecasting techniques into

- Subjective and objective forecasting techniques, and
- Causal and non-causal forecasting techniques.

The various models described below as per (Pycraft *et al*., 2003) can be depicted as follows as per Figure 3:

	Non-Causal Techniques	**Causal Techniques**
Objective Techniques	**Time Series Analysis:** • Moving-average Smoothing • Exponential Smoothing	**Regression** **Economic Models**
Subjective Techniques	**Intuition**	Individual Expert Opinion Group Expert Opinion (e.g. Delphi forecasting)

Figure 3: Classification of forecasting techniques

They indicate that subjective forecasting are those that involve "judgement and intuition from one or more individuals, whose approach to the forecasting task is unlikely to be explicit, but will be based on experience" whilst objective techniques have "specified and systematic procedures". Causal techniques according to them use the basis of causal relationships in the predictions i.e. if the cause-effect relationship between variables can be modelled then predictions of the factors" that influence the item of forecast will enable such forecast. The assumption of these methods is that causal variable can be measured and projected more accurately than the actual variable itself. Non-causal techniques, as per Figure 3 are those that "use past values of a variable to predict future values" (Pycraft *et al*., 2003: 800). The assumption of these models is that underlying causes of events of the past, "will continue to shape events in exactly the same way in the future" (Pycraft *et al*., 2003 : 800).

(Makridakis *et al*., 1998) classify the forecasting techniques in two major categories:

- Quantitative methods for which sufficient quantitative data is available and

- Qualitative for which little or no quantitative data exists but there is sufficient qualitative knowledge available.

Quantitative methods include time series analysis based on historical data and explanatory methods that are based on "understanding how explanatory variables" affect the variable being forecast. (Makridakis *et al*., 1998: 9) also state that there are three conditions that must exist to apply quantitative methods:

- "Information about the past is available"
- "The 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". This is also the assumption of continuity and is the underlying requirement of all quantitative and qualitative forecasting methods.

These conditions are also discussed by (Saffo, 2007). He indicates that a historical "rearview mirror is an extraordinarily powerful forecasting tool. The texture of past events can be used to connect the dots of present indicators and thus reliably map the future's trajectory – provided one looks back far enough". According to (Saffo, 2007) "The recent past is rarely a reliable indicator of the future" and advocates that one "look back at least twice as far as you are looking forward".

(Gaither and Frazier, 2002) also categorises forecasting methods into qualitative and quantitative forecasting models. They indicate that qualitative methods are based on judgements of the "causal factors that underlie" the item being forecasted and the "likelihood of those causal factors being present in the future". Qualitative methods can involve several levels of sophistication from scientific opinion surveys to "intuitive hunches about future events" (Gaither and Frazier, 2002: 66)). They also classify quantitative models as "mathematical models based on historical data" that assume that "past data is relevant to the future" (Gaither and Frazier, 2002: 67).

(Georgoff and Murdick, 1986) list forecasting methods as follows:

- **<u>Judgement Methods</u>**:
    - **Naïve extrapolation** that is an extension of the results of current events;
    - **Sales-Force Composite** that is a composite of estimates by sales staff that is adjusted for biases and expected changes;
    - **Jury of Executive Opinion** that is based on consensus of group of "experts" for different functional areas in a company;
    - **Scenario methods** that present unfolding narratives about events in the future;
    - **Delphi methods** that uses successive series of estimates by a group of experts that iteratively uses the group's previous results for new estimates;
    - **Historical analogy** that provides predictions on elements of the past events that is analogous to the present.
- **<u>Counting Methods</u>**:
    - **Market testing** through representative buyers or responses to new products or offerings extrapolated to estimate future prospects of product or service;
    - **Consumer market survey** based on data gathered from consumers on attitudes and purchase intentions;
    - **Industrial market survey** similar to the consumer market survey but sampling of fewer knowledgeable subjects.
- **<u>Time Series Methods</u>**
    - **Moving averages** that use averages of the recent values of the forecast variable to make predictions of the future;
    - **Exponential Smoothing** that uses a constantly weighted combination of the forecast estimate for the previous period and the most recent outcome;

  - **Adaptive filtering** that is a derivation of a "weighted combination of actual and estimated outcomes, systematically altered to reflect data pattern changes"; (Georgoff and Murdick, 1986);
  - **Time Series Extrapolation** where predictions are derived from the future extension of least squares function fitted to a data series using time as an independent variable;
  - **Time series decomposition** that bases forecasts on an analysis of the expected outcomes from trend, seasonal, cyclical and random components of the data series;
  - **Box-Jenkins** model that is a complex iterative procedure that is based on an autoregressive, integrated moving average model adjusted for seasonal and trend factors and estimates the appropriate weighting parameter. This model is part of **Autoregressive methods** such as Autoregressive Moving Average (ARMA) models and Autoregressive Integrated Moving Average (ARIMA) models (Meade, 2000). (Wang, 2008) indicates that for the Box-Jenkins model the I in the ARIMA model indicates that the "time series has been transformed into a stationary time series".
- **<u>Association or causal methods</u>**:
  - **Correlation methods** that bases forecasts on historic patterns of covariation between variables;
  - **Regression models** that minimizes the residual variance of one or more predictor (independent) variables from a predictive equation;
  - **Leading indicators** that generate forecasts from "one or more preceding variables that are systematically related to the variable" (Georgoff and Murdick, 1986) to be forecasted;
  - **Econometric models** that use "an integrated system of simultaneous equations that represent relationships among

elements of the national economy derived from combining history and economic theory" (Georgoff and Murdick, 1986) and

- **Input-output models** that indicate demand changes of one industry that directly and cumulatively affect other industries through a matrix model.

The objective of the research proposal is to find quantitative methods for forecasting in the commuter rail environment and therefore two methods discussed above that are of relevance to the study objective are time series analysis and association or causal statistical techniques.

2.2.1 Time Series Analysis:

Time series analysis according to (Foster, Barkus and Yavorsky, 2006) examines the trends between repeated observations taken over equal time intervals and addresses changes of trends, seasonal changes and random fluctuations, of the observations over time.

The dependent variable in time series analysis consists of observations taken over time on a number of occasions and addresses research questions of:

- Are there discernable patterns in the data over time;
- Is the dependent variable impacted in by an independent variable measure over the time series?
- Is there an equation that can be derived that will forecast the future observations of the dependent variable effectively? (Foster *et al.,* 2006).

The types of time series analysis are as follow according to (Foster *et al.*, 2006):

- Simple time series design: This design is the most common and consists of quantitative measures taken over regular time periods.

- Cohort analysis design: This design involves the study of a group of people that experienced a particular significant event roughly at the same time and a follow up study where individuals are examined again must be representative of the original group.
- Panel studies design: In this design the same people are followed over time.
- Event history design: This design uses events that occur over time and usually refers to organisational behaviour rather than individuals.

The simple time series analysis is applicable to the study objective and the other designs will not be considered.

(Kedem and Fokianos, 2002:4 and 5) provides the mathematical formula for the explanatory variables or covariates for a time series $\{Y_t\}$ with the aim to predict or forecast as

$$\boxed{Z_{t-1}=(Z_{(t-1)1},\ldots..,Z_{(t-1)p})'}$$

Where p = dimensions and

t = 1, ….., N

and the expression of the past as

$$\boxed{\mathcal{F}_{t-1} = \sigma\{Y_{t-1},Y_{t-2},\ldots....,\mathbf{Z}_{t-1},\mathbf{Z}_{t-2}, \ldots..\}}$$

Different types of time series analysis are according to (Foster *et al.,* 2006), the Spectral or Fourier analysis, time domain analysis and forecasting. In Spectral of Fourier analysis "converts the plot of the dependent variable over time into its sine or cosine wave components to produce a sum or integral of these functions and examines the degree to which the plot differs from the sine and cosine functions" (Foster *et al.,* 2006: 117). This type of time series analysis is normally used in mathematics and theoretical physics and due to the inability of statistical packages to analyse multivariate time series analysis its use is restricted. (Foster *et al.,* 2006). Time domain analysis is mostly used in social sciences and uses the raw dependent variable directly as opposed to the conversion of spectral analysis. Forecasting is a way in which the time series analysis are used and often gets used in conjunction with regression analysis. (Foster *et al.,* 2006).

Time series analysis aim is to express the underlying trends in the data as equations that form the model of the data and will be used for forecasting of the data based on the model. (Foster *et al.,* 2006). These models generated by the underlying trends can either be linear or non-linear. The linear models include auto-regressive (AR), moving average (MA) and combinations of these (ARMA). Non-linear models include exponential auto-regressive (EXPAR), threshold auto-regressive (TAR) and auto-regressive conditional heteroscedastic (ARCH) models. The most frequent used model is the auto-regressive, integrated, moving average model (ARIMA) or Box-Jenkins model. (Foster *et al.,* 2006).

(Hill and Lewicki, 2006) indicates that the Box-Jenkins or ARIMA model, includes the following components:

- Autoregressive i.e. each observation consists of a random error component and a linear combination of prior observations,
- Moving average components and
- Explicitly differencing on the formulation of the model to achieve the stationary requirement of the model.

(Van den Bergh, Holloway, Pienaar, Koen, Elphinstone and Woodborne, 2008) indicates that ARIMA models "are a subset of time series analysis techniques that may be used to forecast future values of a time series based on historical values of a time series". Seasonality is accommodated by ARIMA models as well as local seasonality. Local seasonality is "data more related to the same season one or two years previously than the same season several years ago" (Van den Bergh *et al*., 2008). They further indicate that autocorrelation in ARIMA models can be used to analyse the time-dependency of data. (Van den Bergh *et al.,* 2008) also indicates that univariate ARIMA models can be powerful for short forecast horizons, but not be appropriate for forecasting the longer term.

The time series models focused on forecasting in line with the study objective will be considered.

2.2.2 Association or causal statistical techniques:

The general linear model (GLM) is the statistical theory that underpins many parametric statistical techniques (Foster *et al.,* 2006) and is fundamental according to (Timm, 2002) to the analysis of univariate and multivariate data. The GLM method is used "to determine whether the independent variable(s) affect or relate to the dependent variable(s)" (Foster *et al.,* 2006: 11) and according to (Kedem and Fokianos, 2002:xiii) "provides under some conditions a unified regression theory suitable for continuous, categorical and count data". GLM also successfully addressed non-normal observations found in binary and count data according to (Kedem and Fokianos, 2002). The following regression equation, that indicates that the dependent variable Y is related to the independent variable X, depicts the GLM:

$$\boxed{\mathrm{Y} = c + b\mathrm{X}}$$

(Foster *et al.,* 2006:11) or as per (Timm, 2002: 106)

$$\boxed{Y_{N\ x1} = X_{Nxk}\beta_{kx1} + e_{Nx1}}$$

Where

k = number of parameters or independent variables

$\beta =$ relationship between Y and X and

e = random error with a mean of 0 and

the covariance is Ω

This equation assumes that the relationship between the variables is linear in nature. Variables used in the statistical techniques are assumed to have an additive effect that means they are contributing to the prediction or forecast of the dependent variable (Foster *et al.,* 2006). (Foster *et al.,* 2006) also indicates that the GLM underlies the following statistical techniques:

- Analysis of Variance (ANOVA) where differences of three or more groups are examined.
- Analysis of Covariance (ANCOVA), an extension of ANOVA where a covariate or another relevant variable is used as an additional independent variable. This variable can be controlled by the researcher.
- Multivariate analysis of variance (MANOVA) and Multivariate analysis of covariance (MANCOVA) an extension of ANOVA for multiple continuous dependent variables in the analysis.
- Regression that involves correlations or relationships between pairs of variables and multiple regressions where a single dependent variable is predicted or forecasted from a number of independent variables.
- Log-linear Analysis that are used to analyse contingency tables (for two variables) or cross-tabulations for more than two variables.
- Logistic regression is used to forecast one dichotomous dependent variable from one or more predicting variables.

- Factor analysis or principle component analysis normally from questionnaire items, tries to reduce a large number of variables to a few factors or components
- Structural equation modelling or causal modelling that deals with multiple independent and dependent variables of categorical or continuous data and
- Survival or failure analysis used in medicine or component failure. This method considers many independent variables dependent on the research question.

(Van den Bergh *et al.,* 2008) indicates that GLM's allow the dependent variable to follow a number of distributions namely normal, Poisson, binomial, exponential and gamma distributions. As ticket sales are basically count data that are not negative a GLM might be appropriate.

With GLM's (Van den Bergh *et al*.,2008) indicates that "the choice of distribution to fit to the dependent variable is important". The following criteria for choice of a distribution are indicated by (Van den Bergh *et al.* 2008):

- For count data not in the form of proportions and a variance equal to the mean the Poisson distribution can be appropriate;
- If the variance is much larger than the mean a negative binomial distribution may be more appropriate.

The methods of ANOVA, MANOVA as well as multiple regression methods will be most applicable to the study objective of this research in conjunction with time series analysis.

2.2.2.1 ANOVA model

(Timm, 2002) also provide the univariate regression model as

$$y_{nx1} = X_{nxq}\beta_{qx1} + e_{nx1}$$
$$\mathrm{cov}(y) = \sigma^2 I_n$$

Where q = k+1

k = number of independent variables

e = random errors with mean zero and common unknown variance σ^2

The ANOVA model is subset of the univariate regression model when there is less independent variables than q and X_i are indicator variables.

ANOVA consider the effect of more than one independent variable and whether the variables interact i.e. the effect of one influenced by the other and the amount of variance in the data set (Foster *et al.,* 2006). The F-value of ANOVA test is the ratio of the mean square between-groups and mean square within groups and the further away the F-value from 1, the less the probability that differences between groups are based on chance and thus higher the significance of the effect of the independent variable (Foster *et al.,* 2006).

2.2.2.2 MANOVA model:

The purpose according to (Foster *et al.,* 2006:16) of the MANOVA model is to "determine whether multiple levels of independent variables on their own or in combination with one another have an effect on the dependent variable". (Timm, 2006) indicates that the MANOVA model is a special case of linear multivariate regression that uses the same set of independent variables, X, to model the set of dependent variables Y.

(Timm, 2002:111) provides the multivariate linear regression model as :

$$Y_{nxp} = X_{nxq}\,\beta_{qxp} + E_{nxp}$$
$$=[X\beta_1, X\beta_{2,}.......,X\beta_p] + [e_1,e_2,.....,e_p]$$

Where n = number of observations,

k = number of independent variables

q = k + 1

p = number of correlated dependent variables and

e = random errors

The MANOVA model is a subset of multivariate linear regression model for each p correlated dependent response variables when there are less independent variables then q and X_i are indicator variables.

2.2.2.3 Multiple Regression:

(Diamantopoulos and Schlegelmilch, 2000:214) indicate that multiple regression is used to "analyse the relationship between one dependent variable and a number of independent variables" and that both dependent and independent variables need to be measured on interval or ratio level. (Foster *et al.,* 2006) indicates that this statistical model is used to assess the relative influence if a number of predicting variables are used to forecast a dependent variable and which of these predicting variables are more important. This method requires the dependent variable on a continuous scale and the measures or values of two or more predicting variables that can be on a continuous scale, categorical values or a mixture of these. The method only provides:

- the relative importance of predictors provided and not the important predictors as this is a function of the researcher,
- the relative importance of the predictors as related to the values provided and it cannot be extrapolated to a different set of observations; and

- the method does not establish causation

The regression equation for multiple regression models as per (Foster *et al.,* 2006: 34) is:

$$Y = a + B_1(x_1) + B_2(x_2) + \ldots + B_k(x_k)$$

Where B1, B2, etc are regression coefficients or weights and
X1, x2 etc are the independent variables.

(Steel and Uys, 2007) indicates that one of the first steps in a multiple linear regression analysis is the statistical variable selection. They state that "the purpose of regression variable selection is to reduce the predictors to some "optimal" subset of the available regressors" (Steel and Uys,2007). They also state that reduction of predictors is required as

- A smaller set of variables can provide more accurate forecasts or
- Predictor variables that significantly influence the response can be identified.

According to (Steel and Uys, 2007) techniques that can be used for variable selection are stepwise routines, Bayesian techniques, cross-validation selection techniques or an all possible subsets approach. Their selection of all possible subsets approach is based on Mallows C_p criteria or Akaike's information criteria (AIC). The aim of the selection based on the above criteria is that the presence of the predictor case should improve the fit of the selected model.

2.2.2.4 Regression models and time series analysis

The use of partial likelihood is used to transport the main "inferential features appropriate for independent data to time series … following generalised linear models" (Kedem and Fokianos, 2002:1)

Time series data following GLM need to follow the following stipulations according to (Kedem and Fokianos, 2002) namely the

- Random component that requires that the distribution of the past values are part of the exponential family of distributions and
- Systematic component that for *t* = 1, ….,N there is a monotone link function

$$g(\mu_t) = \eta_t = {}_{j=1}\Sigma^p \beta_j Z_{(t-1)j} = Z'_{t-1}\beta$$

Where η_t is the "linear predictor of the model" (Kedem and Fokianos, 2002:6)

with a typical choice for $Z_{t-1}\beta$

$$Z_{t-1}\beta = \beta_o + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-2} X_t + \beta_4 Y_{t-2} X_{t-1}$$

Observations that are non-negative and integer-valued over time are also known as count time series. (Kedem and Fokianos, 2002). The data for commuter rail sales (ticket issues) are also count time series. Modelling of dependent counts time series can be done successfully through the Poisson distribution. (Kedem and Fakianos, 2002).

2.2.2.4.1 Time Series Poisson model:

The density of the Poisson distribution for application with time series with count data with mean μ_t can be expressed as

$$f(y_t; \theta_t, | \mathcal{F}_{t-1}) = \exp\{(y_t \log\mu_t - \mu_t) - \log y_t!\},\ t = 1,\ldots,N$$

so that $E[Y_t| \mathcal{F}_{t-1}] = \mu_t, b(\theta_t) = \mu_t = \exp(\theta_t), V(\mu_t) = \mu_t, \emptyset = 1$ and $\omega_t = 1$

The canonical link is

$$g(\mu_t) = \theta_t(\mu_t) = log\mu_t = \eta_t = Z'_{t-1}\beta$$

(Kedem and Fokianos, 2002:9)

The Poisson distribution of past values can also be expressed as

$$f\ (y_t;\mu_t|\mathcal{F}_{t-1}) = (\mathrm{Exp}(-\mu_t)\mu_t^{y^t})/\ y_t!,\ \ t = 1,\ldots..,N$$

(Kedem and Fokianos, 2002:140)

For this Poisson model its required variance according to (Kedem and Fakianos, 2002) is:

$$E[Y_t\ |\mathcal{F}_{t-1}] = \mathrm{Var}\ [Y_t\ |\mathcal{F}_{t-1}] = \mu_t,\ t= 1,\ldots,N.$$

and $\{\mathbf{Z}_{t-1}\}$ for t= 1,….,N a *p*-dimensional covariate that can include past values of the process and/or additional auxiliary information. A typical choice for $\mathbf{Z}_{t-1}$ is also according to (Kedem and Fakianos, 2002):

$$\mathbf{Z}_{t-1} = (1,\ X_t,\ Y_{t-1})'$$

Where $X_{t\ i}$is an additional process.

This leads according to (Kedem and Fokianos, 2002) to a suitable model for analysis of count time series as :

$$\begin{aligned}\mu_t(\beta) &= h(\mathbf{Z}'_{t-1}\beta),\ t= 1,\ldots..,N\\ &= \exp(\eta_t)\\ &= \exp(\mathbf{Z}'_{t-1}\beta) \text{ or log-linear model.}\end{aligned}$$

Where β is a p-dimensional vector of unknown parameters and the $h(.)$ the inverse link function.

The partial likelihood estimation of the Poisson Model for β according to (Kedem and Fakianos, 2002) is

$$\begin{aligned}PL(\beta) &= {}_{t=1}\Pi^N f(y_t;\boldsymbol{\beta}|\ \mathcal{F}_{t-1})\\ &= {}_{t=1}\Pi^N\ (\text{Exp}(-\mu_t)\mu_t y_t/y_t!) \text{ for } t = 1,\ldots..,N\end{aligned}$$

and the partial log-likelihood is

$$\begin{aligned}l(\beta) &= \log PL(\beta)\\ &= {}_{t=1}\Sigma^N\ y_t \log\mu_t(\beta) - {}_{t=1}\Sigma^N\ \mu_t(\beta) - {}_{t=1}\Sigma^N \log(y_t!)\end{aligned}$$

the partial score function through differentiation is

$$\mathbf{S}^N(\beta) = {}_{t=1}\Sigma^N\mathbf{Z}_{t-1}(\partial h(\eta_t)/\partial\eta_t)(1/\sigma^2{}_t(\beta))(Y_t-\mu_t(\boldsymbol{\beta}))$$

Where $\eta_t = \mathbf{Z}'_{t-1}\beta$ and $\sigma^2{}_t(\beta) = \text{Var}\ [Y_t\ |\mathcal{F}_{t-1}]$

The score equation that provides the maximum partial likelihood estimator is

$$\mathbf{S}_N(\beta) = \nabla\log\ (PL(\beta) = 0$$

a non-linear system that can be solved by the Fisher scoring method.

2.2.2.4.2 Time Series Autoregressive Conditionally Heteroscedatic (Arch) Models

Heteroscedasticity according to (Hanke and Reitsch, 1998: 259) "exists when errors or residuals do not have a constant variance across an entire range of values". (Bickel, 2007) indicates that there are various correctives for heteroscedasticity and that the easiest of these methods is the estimated generalized least squares method. (Bickel, 2007) also provides a test for heteroscedasticity namely the Koenker-Basset (KB) test. Alternatively for time series exhibiting large variability or volatility autoregressive conditionally heteroscedastic (ARCH) models may be used. The ARCH model according to (Kedem and Fokianos, 2002) can be specified as:

$$Y_t = \sigma_t \text{ and}$$
$$\sigma^2_t = \beta_0 + \beta_1 Y^2_{t-1}$$

Where coefficient β_1 is assumed to be positive and $\{\in_t\}$ is a sequence of standard random normal variables.

The conditional distribution of Y_t is a normal distribution with mean equal to 0 and variance $\beta_0 + \beta_1 Y^2_{t-1} + \mu_{t.}$ This can then be converted to reflect a scaled $\mathcal{X}^2$ random variable namely:

$$Y^2_t = \beta_0 + \beta_1 Y^2_{t-1} + \mu_t$$

This then follows an autoregressive process according to (Kedem and Fokianos, 2002) for $0 \le \beta_1 < 1, 3\beta^2_1 < 1$ and for a finite variance of $\{\mu_t\}$. The equations for this are

$$E[Y^2{}_t] = Var\ [Y_t] = \beta_0\ /(1\text{-}\beta_1)\ \text{and}$$

$$Var\ [Y^2{}_t] = E[Y_t{}^4] = 3\beta^2{}_0/(1\text{-}\beta_1)^2(1\text{-}\beta^2{}_1)/(1\text{-}3\ \beta^2{}_1)$$

The kurtosis of this marginal distribution is equal to

$$E[Y_t{}^4] / [E[Y^2{}_t]]^2 = 3((1\text{-}\beta_1)^2/(1\text{-}3\ \beta^2{}_1)$$

and if 3 $\beta^2{}_1 \geq 1$ then $\{Y^2{}_t\}$ is stationary with infinite variance.

The estimate for the parameter vector $\beta = (\beta_0,\beta_1)$ is done through the maximisation of the log-likelihood of

$$l(\beta) = \text{-}½_{t=1}\Sigma^N\{\log(\beta_0 + \beta_1 y^2{}_{t-1}) + (\ y^2{}_t/(\ \beta_0 + \beta_1 y^2{}_{t-1})\}$$

2.2.2.4.3 Model for single-season heteroscedaticity

(Tripodis and Penzer, 2007), found that for time series where the variance of one season differs with others, adjustments with a dummy seasonal variable or trigonometric seasonality is not effective in modelling a series with single-season heteroscedaticity. They propose that an approach is used where periodic heteroscedaticity measurement noise is superimposed on homoscedastic seasonality. This is similar to a deseasonalized model where the seasonal component is deterministic and the variance of the irregular component depends on a single season. The periodic heteroscedaticity can be reflected in a periodic seasonal variance model or a periodic irregular variance model.

The time series seasonal differences is defined as :

$$z_t = \Delta_s\gamma_t = \lambda_t + \Delta_s\varepsilon_t$$

Where

γ is the seasonal component of the time series and

ε is the irregular or white noise component

The autocovariance function for the seasonal differences is:

$$= \operatorname{cov}(z_t, z_{t-h}) = \begin{cases} c_\lambda(r,0) + 2\sigma^2_{\varepsilon,r} & h=0 \\ c_\lambda(r,h) & h=1,\ldots.,s-1 \\ -\sigma^2_{\varepsilon,r} & h=s \end{cases}$$

with $c_z(r,h) = 0$ for $h>s$.

For the periodic seasonal variance model,

$$\sigma^2_{\varepsilon,r} = \sigma^2_\varepsilon$$

for all r i.e. season-dependent at lags $h = 0,\ldots\ldots,s-1$.

For the periodic irregular variance model

$$c_\lambda(r,h) = c_\lambda(1,h)$$

for all r i.e. periodicity is restricted to the variance and lag s covariance.

The autocovariance function is used as an indicator of the periodic behaviour. In a periodic irregular variance model the relation between seasons is the same within each year, whereas in the periodic season variance model, “the relation between the unusual season and all other seasons differs from the relation between any other two seasons” (Tripodis and Penzer, 2007).

(Tripodis and Penzer, 2007) further indicates that “there are several models to test for periodicity in the autocorrelation function” but that the power of these tests are very small for samples of less than 30 years rendering them impractical for economic time series. They also indicate that few tests are available to test for seasonal heteroscedaticity although seasonal heteroscedaticity is relatively common. “Existing tests are based on likelihood ratio, Wald or Lagrange multiplier principles” (Tripodis and Penzer, 2007). They further indicate that useful information on seasonal heteroscedaticity can be obtained graphically of time series plots, month plots and correlograms.

2.3 Usage of different forecasting models:

(Makridakis *et al*., 1998) provide a summary of surveys among forecasting users between 1984 and 1996 on the familiarity and satisfaction levels with major forecasting methods. Judgmental forecast models were most familiar to forecasting users followed by simpler quantitative methods of moving averages, straight-line projections and exponential smoothing as well as regression methods. Box Jenkins methods applied on ARIMA models were the least familiar to forecasting users and time series classical decomposition the second least familiar method. On the issue of satisfaction levels regression methods rated the highest satisfaction levels followed by exponential smoothing. The methods of moving averages and trend-line analysis also rated high on satisfaction levels. Classical time series decomposition and expert systems and neural networks also rated low on satisfaction. The method with the lowest satisfaction levels in the 90’s was the Box-Jenkins method.

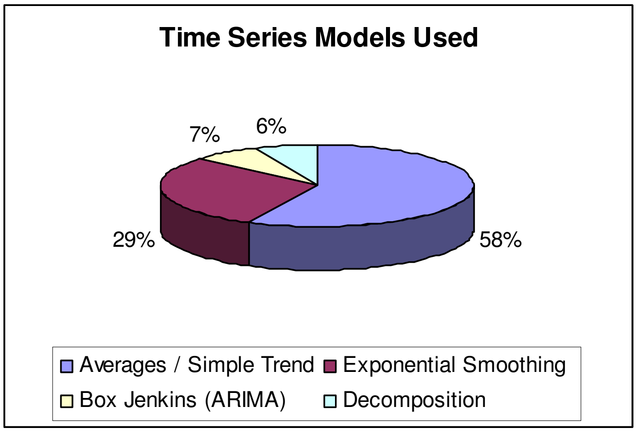
(Price and Sharp, 1988) also reported that growth curve models, widely used, lack robustness, indicating the need for caution in using more complex models. (Makridakis and Hibon, 1997) also found that AR-models can produce more accurate post-sample forecasts than the use of Box-Jenkins model and that “simple methods such as exponential smoothing outperform, on average the Box-Jenkins methodology to ARMA models”.

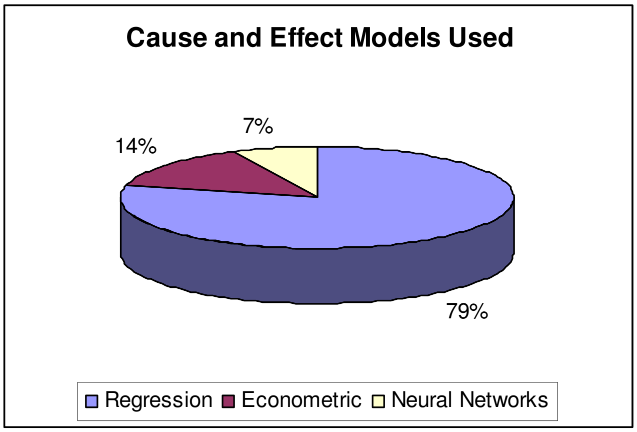
(Meade, 2000) indicates that “the characteristics of the data series are an important factor in determining relative performance between methods” and that “statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones”. (Meade, 2000) propose that using summary statistics such as standard error, median and mean is a method to assist in selection of the most appropriate forecasting method and evaluated this for

- Naïve methods: long-run average, using the last observation as forecast, deterministic trend,
- Linear trend models: single exponential smoothing, Holt’s linear trend, Holt’s damped trend, Robust Trend, ARMA and ARIMA models.

(Wang, 2008) indicates that the tools for identifying a good tool for stationary time series with the ARIMA model are the autocorrelation function and the partial autocorrelation function. By plotting these functions on a correlogram assists in choosing the correct ARIMA model.

(Jain, 2008) found that time series models (61%) were the most used in business in the USA, followed by Cause-and-effect models at 18% and lastly judgemental models at 15%. Five percent of companies used their own homegrown models. Where there is more uncertainty (Jain, 2008) found that judgmental models were higher in use. The usage of the time series and cause and effect models in the United States of America were as follow based on a 2007 survey:





Some of the other more complex models developed by academics are as follow:

- Usage of seasonal models such as Harrison and Stevens model incorporating seasonal heteroscedasticity (when one or more seasons are more variable than others) through trigonometric seasonal models. (Proietti, 1998));
- Multivariate seasonal growth multiplicative model that consists of a linear trend component for each individual series (Sadownik and Barbosa, 1999)
- Non-linear method of analysis based on an econometric model and forecasting for vector time series with a multiplicative seasonal component (Sadownik and Barbosa, 1999) and;

- Fractionally integrated autoregressive moving-average (FARIMA) model with k-factor Gegenbauer Processes. (Ferrara and Guègan, 2001) used in forecasting urban transport traffic in Paris.

(Hanke and Reitsch. 1998) provides the following guidelines for models to be used for forecasting series:

- Forecasting data with a trend: use linear moving average or Holt's linear exponential smoothing or simple regression or the Gompertz model or exponential models.
- Forecasting data with seasonality: use classical decomposition, Winter's exponential smoothing, time series multiple regression or Box-Jenkins methods.
- Forecasting cyclical series data: use classical decomposition or economic indicators or econometric models, multiple regression and / or Box-Jenkins method.

2.3.1 Forecasting accuracy and bias:

(Landram, Pavur and Alidaee, 2008) indicates that although accurate forecasts are important to business, each business need to select the forecasting method that help their particular situation best. They further indicate that forecasting is complicated by continuously changing economic conditions. (York, 2005) indicates that "the only thing that can be said with near certainty about a forecast is that it will be wrong". This indicates that the reasons for inaccuracies need to be explored as well as that methods to improve the accuracy are required.

The analysis of (Makridakis *et al*., 1998) found that accuracy is not the only method for selecting a forecasting method. (Flores and Wichern, 2005) state that "If a global measure of forecast accuracy is required …. the "metric used should be relatively simple, easily described and displayed, and be useful in a sense that an individual deviation from target, or changes in the measurement over time, provide relevant information for decision-making". They suggest the use of a simple graphical boxplot to monitor the forecast over time.

Judgment methods are subject to biases and limitations and have been found to on average more inferior to statistical methods. Memory recall in terms of what is important and what are trivial are the most important limitation of judgmental methods and has been empirically confirmed. Some of the biases of judgmental forecasts are inconsistency, conservatism, recency (more recent events dominate), availability, anchoring, illusory correlations, search for supportive evidence, regression effects, attribution of success and failure, optimism, underestimating uncertainty and selective perception. However the combination of structured judgmental modifications for increasing the accuracy of out-of-sample predictions is promoted by (Landram, et.al, 2008) as influences of the past does not necessarily continue in the future.

Econometric methods were found to be better than time series models but more complex models did not perform better than simpler ones. Multivariate models (ARIMA, Vector autoregressive) did also not show superior performance. Non-linear models also did not provide better accuracy. Macroeconomic forecasts based on extrapolation of historical trends were found to be mostly inaccurate. (Tay and Wallis, 2000) reported that in macroeconomic forecasting, forecast intervals are used. One can see this as a reaction to the reported inaccuracy of a point forecast. Empirical evidence did not support the better forecasting ability of adaptive parameters versus fixed parameters. Lastly expert systems and neural networks also did not perform better than exponential smoothing. (Makridakis *et al*., 1998)

Simple methods were found to be just as good as complex or statistically sophisticated ones for post-sample forecasts using a few series and judgmental adjustments are made by experts in the forecasting (Makridakis *et al*., 1998). (Fader and Hardie, 2005) found that through the process of building a simple model, the modelling environment and data management constraints are considered allowing modification of the data structures and either allow simplification of the mathematical model or else for the development of more complex data structures. They also state that a simple model do sometimes "come at the cost of technical precision" (Fader and Hardie, 2005).

(York, 2005) indicated that inaccuracy of forecasts applies most certainly to a point forecast and that interval forecasts also are susceptible to inaccuracy. Another way to deal with accuracy of forecasts is density forecasts. (Tay and Wallis, 2000: 235) state that a "density forecast of the realization of a random variable at some future time is an estimate of the probability distribution of the possible future values of that variable." The probability distributions used are the Normal distribution, Gaussian distribution and Student's *t*-distribution. Such forecasts provide "a complete description of the uncertainty associated with the prediction, and stand in contrast to a point forecast, which by itself contains no description of the associated uncertainty" (Tay and Wallis, 2000: 235). An intermediate solution is the forecast interval that states the probability that the actual value will fall within the stated interval. (Saffo, 2007) labels this a "cone of uncertainty" that "delineates possibilities that extend out from a particular moment or event". The forecaster's job is according to (Saffo, 2007) to "define the cone in a manner that helps the decision maker exercise strategic judgment".

(Meade, 2000) found that the "more specific the underlying model of the forecasting method, the easier it is to predict its relative performance" and the most effective method is the one with the lowest estimated transformed performance index, $f(v_{ij})$. He uses the values of the β coefficients in conjunction with the values of the summary statistics to predict the performance index of each forecasting method.

A combination of judgment and statistical forecasts to exploit both the methodologies advantages and avoiding their drawbacks are required for use in practice (Makridakis *et al*., 1998).

It was found in empirical studies that the averaging of forecasts of one or more methods results in more accurate forecasts than the individual methods separately and in addition the size of forecasting errors are smaller in this method than each of single methods being averaged. Combination forecasts also reduce variance of the post-sample forecasting errors. (Landram *et al.*, 2008) indicates that the manner in which forecasts are best combined is still being explored. They combine forecasts of time series components with forecasts of least-squares modelling. This modelling approach reduces the error sum of squares (SSE). The combination of time series components of trend, seasonal and cyclical with other forecasts was found by (Landram *et al.*, 2008) to make a significant contribution in explaining the dependent variable. One of the other methods of combination is the combination of additive and multiplicative forecasts when the accuracy of multiplicative model approximates that of the additive model. The combination per (Landram *et al.*, 2008) is formulated as follow:

$$\hat{Y}_t = b_0 + b_1X_t + b_2S_j + b_3C_t + b_4T_tS_jC_t$$

Where

X_t represents the time periods

S_j is the quarterly seasonal indices and

C_t is the cyclical indices.

(Landram *et al*., 2008) indicates that the intercept value should be included for combination of forecasts with least square models with automatic assignment of weights to the forecast, to improve the forecast accuracy.

Factors that reduces the accuracy of individual forecasts (Makridakis *et al*., 1998) are

- Measuring the wrong variable based on data that are available such as orders or shipments in stead of actual demand.
- Measurement of errors due to clerical, data processing errors, accounting changes and data definition changes.
- Unstable or changing patterns or relationships as these patterns are not constant as assumed by statistical models. Systemic changes introduce non-random errors in forecasting.
- Models that minimise one-period-ahead errors are not suitable for several or many periods ahead.

(Gaither and Frazier, 1999: 92) states that reasons for ineffective forecasting are due to

- Failure to involve a "broad cross section of people in forecasting";
- "Failure to recognise that forecasting is integral to business planning";
- "Failure to recognise that forecasts will always be wrong." This is due to the error and the fact that the magnitude of error tends to get bigger over long time periods.
- "Failure to forecast the right things";
- "Failure to select an appropriate forecasting method" and
- "Failure to track the performance of the forecasting model"

2.4 Forecasting method selection:

Statistical forecasting methods allow forecasters to extrapolate established patterns and existing relationships to forecast their continuation based on the assumption that the patterns and or relationships will not change during the forecasting phase. The detection of changes or knowledge of changes that will occur requires human judgement that is the only viable alternative for prediction of the change and the impact and extent of such changes.

(Foster *et al.,* 2006) indicates that data predicted from a time series model must be based on a model that explains the data reliably. They also state that future data can only be predicted if the future circumstances are similar to those which existed when the original data was collected and therefore the forecasts from time series must be treated with caution and constraints resulting from the initial data be stated explicitly.

(Makridakis *et al*., 1998) states that an appropriate method for forecasting should be based on whether forecasts or explanations are required, the method's accuracy, the data, characteristics and type of the data, the range and frequency of the forecast.

(Fader and Hardie, 2005) states that ease of implementation from the forecast user's side, needs to be considered for the usage of the models developed by academics. Increasingly complex models are not understood by managers that do not have the mathematical or statistical training and if understood by the manager the models need to be explained and sold to others in the organisation that have limited understanding of the models and that most models require relatively sophisticated modelling skills with "custom programming and non-standard data manipulation" (Fader and Hardie, 2005) that have large funding commitments and high risk of failure. They also state that "fewer and fewer companies have specialised departments in which such skills reside" (Fader and Hardie, 2005). This leads to the notion of a simple but not naïve model(s) that are developed in an evolutionary process to more complex models to "foster managerial acceptance, encourage an orderly development of data and analysis systems, and reduce risk of failure" (Fader and Hardie, 2005).

(Jain, 2008) indicates that the fundamentals in model selection are:

- Actual = Pattern + Error. Thus each dataset forms a specific data pattern and the model captures a specific data pattern. The objective is to find a model that captures most of the pattern inherent in the data.

- 100% Accuracy is not possible or necessary. The objective is to minimise the error as much as possible.
- Sophisticated models as indicated are not necessarily better and simpler models are often easier to explain.
- Magic models do not exist and matures with age. Maturity of a model requires a change to the model.
- Each model has its own data requirements
- Statistical forecasts are a baseline forecast and can be improved with judgment overlay and should therefore not be prepared in isolation.

(Flores and Wichern, 2005) also states that there is not one uniformly best forecasting method regardless of the error metric employed. They also found that "if some of the forecasted demands are biased, as is likely the case if a single forecasting method is used, aggregate measures of performance can be misleading". (Flores and Wichern, 2005). The article from (Flores and Wichern, 2005) also mentions that a "good number for accuracy is 70 -75%" and "For bias it is +-20%". Accuracy levels below 60% are troublesome and will affect service levels and inventories.

(Jiang, Au & Tsui, 2007) states that "to model and track thousands of diversified customer behaviours, it is important to develop simple and unified robust modelling tools, which can accommodate different behaviour patterns including business changes". Thus the model must be simple but robust for use.

(Foster *et al.,* 2006) state that for time series analysis there need generally be a minimum of 50 dependent variable observations over time. In addition the dependent variable needs to meet the parametric requirements of homogeneity of variance and be normally distributed. The dependent variable observations taken at different times should be equidistant i.e. hourly, daily, weekly, monthly etc.

The dependent variables for statistical method of MANOVA need to confirm with parametric assumptions of normality and homogeneity of variance i.e. the amount of variance is the same in the different observations. This requirement is however not important for groups of equal sizes (Foster *et al.,* 2006). For complex MANOVA methods, a robust analysis are linked to sample size and large samples are recommended for this method.

For the multiple regression method the number of values should be at least 40 + k where k is the number of independent variables. The data for multiple regression should be examined for outliers, meets the requirement of normal distribution, linear relation and homoscedaticity ("variance of the dependent variable does not differ at different levels of the independent variable" (Foster *et al.,* 2006: 38) It is also a requirement that the independent variables are not undesirably correlated (multicollinearity). (Foster *et al.,* 2006).

Forecasting is often also subject to whether users of forecasts want to better understand factors that influence the variable we want to forecast in order to influence the direction of the forecasted variable. Regression or econometric explanatory models need to be developed to provide this information (Makridakis *et al*., 1998).

(Meade, 2000) indicates three selection options for quantitative forecasting namely:

- Select the best linear trend model for a deterministic trend;
- Select the best method from all methods as per MAPE or
- Select all methods that provide satisfactory forecasts and use an equally weighted combination of their forecasts.

2.4.1 Characteristics of the data:

Seasonality in the data presents fewer challenges for method selection. Classical decomposition is the simplest model for estimating seasonality. For ARMA models, the seasonality must be removed prior to selection of an ARMA model. Holt's model for time series forecasts also are more accurate when seasonally adjusted. The Winter's model makes adjustments for seasonality but the performance is less accurate that the adjusted Holt's model (Makridakis *et al*., 1998).

The level of randomness and the trend-cycle behaviour are key determinants to method selection. For short-term data where randomness dominates the trend-cycles exponential smoothing is the most accurate approach. If trend-cycles dominates the randomness in time series ARMA models are better than smoothing methods to forecast the continuation of the pattern such as cyclicality and trend persistence. Holt's method is preferred when there is little randomness and trend dominates cyclical fluctuations as it assumes the latest smoothed trend can be extrapolated linearly. Damped exponential smoothing is most appropriate when the cyclical component is dominant (Makridakis *et al*., 1998).

(Wang, 2008) indicates that for use of ARIMA modelling, a relatively large sample size is needed to accommodate the loss of data as a result of differencing and lagged structure of the model. The best ARIMA-model according to (Hill and Lewicki, 2006) is the model that produces statistically independent residuals that contain only noise and no serial dependencies in the residuals. In addition the model should be parsimonious.

(Jain, 2008) indicates that it is necessary to analyse data thoroughly in terms of consistency, outliers and structural changes, seasonality, trends, missing values and the number of periods to be used for the forecast.

2.4.2 Type of data:

Yearly, quarterly, monthly, weekly and daily figures are types of data that relates to the characteristics of time series. Randomness diminishes as level of aggregation increases as with yearly data. With averaging 12 months data eliminates randomness in the data whilst trends dominate (Makridakis *et al*., 1998). In daily data randomness dominates while trends are insignificant if present. Quarterly data are in between these two extremes and can also exhibit strong cyclical fluctuations as well as seasonality. Randomness is limited in quarterly data with a stronger trend-cycle indicating that it is less likely that the pattern in the series will change considerably (Makridakis *et al*., 1998).

2.4.3 Range and frequency of forecast:

Daily data requires a greater number of forecasts than monthly data and monthly forecasts requires a greater number of forecasts than for quarterly data and further diminishes when yearly forecasts are made. This results that for yearly forecasts more effort and expenses can be applied whilst simpler more automatic forecasts are more appropriate when doing daily or weekly forecasts. (Makridakis *et al*., 1998).

Short-term forecasts benefit from the inertia in economic and business patterns as well as predictable seasonality and provide usually accurate and reliable forecasts. The larger the numbers of customers or items being forecasted, the smaller the effects of random forces that also provides higher accuracy and reliability of forecasts. Empirical evidence according to (Makridakis *et al*., 1998) suggests that for these statistical, computerised systems show concrete benefits instead of using judgmental forecasts.

Medium-term forecasting can be easy when patterns and relationships do not change. When the time horizon increases so does the chance for changes in the relationship. Medium-term forecasts are mostly used in budgeting and demand prediction of economic and industry variables. (Makridakis *et al*., 1998) found that in general manufacturing firms are more affected by economic cycles than service firms, luxury goods and services are more effective than those producing or servicing necessities; industrial firms are more affected than those in consumer industries and companies in industries with strong competition are more affected that those in industries where competition is less intense. This requires monitoring of critical variables to know the beginning of a recession or boom as soon as possible. Imbalances in an industry are also an indicator of imminent change in terms of correction.

Long-term forecasts are required for capital expansion plans, R&D project selections, launching of new products or services, formulating long term strategies and adaptations for environmental changes. Extrapolation of mega-trends, analogies and scenario-planning are mostly used for these forecasts. The longer the term of forecasting the lesser the accuracy of forecasts as changes can occur to established patterns and relationships. The value of these forecasts lies in generating organisational consensus and establishes the right sense of direction (Makridakis *et al*., 1998).

Jain (2008) found that the forecast horizon depends on the lead-time or how far ahead a decision must be made. Most companies in a survey in 2007 in the USA forecast one year or less ahead. The forecasting data buckets or forecasting presentation namely daily, weekly, monthly or quarterly are equally important and depends on the planning cycle.

(Lapide, 2007) found that most companies in USA increased the frequency of updating demand forecasts. This resulted in benefits in terms of smaller inventories, improving asset utilisation due to increased productivity and facility throughput and higher accuracy of forecasts. These benefits do come at a cost as it requires more planning managers, system changes and other resources. Therefore the benefits must be evaluated against the additional costs for increasing the frequency of forecast updates. (Lapide, 2007) also states that organisation also need to be able to "react fast enough to take advantage of the opportunities" through increased forecasting updates.

2.4.4 Other factors to be considered:

(Makridakis *et al*., 1998) also indicates that other factors to be considered when forecasting is the impact of

- Economic and market forces (e.g. law of demand and supply) and biological laws (e.g. S-shaped increases in growth) that are also discussed by (Saffo, 2007). (Saffo, 2007) states that "The most important developments typically follow the S-curve shape of a power law: Change starts slowly and incrementally, putters along quietly, and then suddenly explodes, eventually tapering off and even dropping back down" and "The art of forecasting is to identify and –curve pattern as it begins to emerge, well ahead of the inflection point."
- "People's preferences, tastes, and budget constraints." (Makridakis *et al*., 1998:574);
- Aspirations to change the future;
- Ability of for example new technologies to change the future;
- The wish to maintain the status quo by some;
- Capabilities of people or organisations to control or slow down change such as use of cartels or monopolies;
- Natural events and the influence of these events on the economic and business environment;
- "Momentum, or inertia, that sustains established patterns and upholds existing relationships" (Makridakis *et al*., 1998 : 574).
- Accidents, strikes, market inefficiencies and psychological factors or mere coincidences and
- The ability to effectively monitor current events and to take corrective action if necessary and possible.

(Saffo, 2007) expands more on the S-curve and feels that forecasters must become "attuned to things that don't fit things people can't classify or will even reject". He also feels that forecasters or decision maker's biggest mistakes come from over-relying on one piece of seemingly strong information because it happens to reinforce the conclusion they have already reached. (Saffo, 2007) feels that "lots of interlocking weak information is vastly more trustworthy than a point or two of strong information".

"Creative insights will always be in short supply, as will the foresight needed to correctly anticipate major future changes, which when exploited can provide huge commercial benefits." (Makridakis *et al*., 1998: 570). Judgmental forecasts must supplement statistical methods through the areas of identification of future forthcoming changes and providing the direction and extent of these changes to ensure adaptation of the statistical models (Makridakis *et al*., 1998).

Chapter three deals with the methodology applied in this research report to obtain a suitable forecasting model for use in the business environment of commuter rail.

3 Methodology:

A combination of steps from (Makridakis *et al*., 1998), (Foster *et al.,* 2006) (Kedem and Fokianos, 2002), (Landram *et al., 2008) and (Meade, 2000)* will be applied for this study namely:

- Gathering the data to enable construction of the model.
- Exploratory analysis of the data including decomposition analysis to determine the strengths of the trend, seasonality and cycles. This step will assist in the selection of a quantitative model. The steps for time series analysis by (Foster *et al.,* 2006) in this phase include identification, estimation and diagnosis. In the identification phase the degree to which the data is stationary will be examined and adjusted through differentiation. In the estimation phase, the parameters are estimated and represent the auto-regressive element. In the diagnosis phase the model is assessed for explanation of all the patterns in the dependent variable.
- Select and fit several quantitative forecast models such as simple time-series models to ARIMA and ARCH models through (Kedem and Fakionos, 2002)'s steps of entertainment of possible sensible models, parameter estimation of the models, hypothesis testing and deviance analysis, examination of residuals, graphical considerations and model selection on information criteria.
- Compare the models in terms of the forecasting errors.
- Assessment of independent intervention variables on the effect of such interventions on the time series.
- Combination of forecasts by overlaying the most applicable forecast with the results of time series decomposition.

Analysis of the data was performed using Microsoft Excel (2003) and Statistica Version 8 of Statsoft Inc. The details of the various steps are discussed in the following paragraphs.

3.1 Gathering of the data:

The data used are statistical or numerical data. The historical data was collected for the period April 2000 – March 2008 for:

- Ticket sales in total,
- Train delays and train cancellations,
- Serious crime incidents reported in the commuter rail system,
- Price Index of fare increases since 2000,
- Consumer price index (CPI) and CPIX (CPI excluding interest rates),
- Fuel prices for petrol, diesel and illuminating paraffin.

3.2 Exploratory analysis:

Exploratory analysis are used to identify any consistent patterns, trends, importance of seasonality, presence of business cycles, any outliers and strength of any relationships. Exploration is conducted through:

- Graphs of the various variables providing a visual picture of the variables;
- Computation of descriptive statistics namely the mean, standard deviation, minimum and maximum values of the dependent variable;

3.3 Selection of quantitative models:

This step involves selecting and fitting of several quantitative forecast models namely simple time-series models, ARIMA-models and multi-regression model. This is based on the sets of assumptions of the parameters of each model. The steps to be following for selection of the best model of the data are:

1. Testing of the time series decomposition model and forecast data using a number of time series models based on a decomposition analysis for the:
    a. Trend;
    b. Seasonality and
    c. Cycle
    d. Forecast the data for 12 months.
2. Apply the Box-Jenkins model for the data through the following steps:
    a. Ensure that the series is stationary;
    b. Transform the series through differentiation should the series not be stationary;
    c. Analyse the autocorrelations and partial correlations;
    d. Experiment with various ARIMA models to find the best model;
    e. Test the model by analysing the residuals and
    f. Forecast the data for 12 months.
3. Apply regression and multi regression modelling. The steps for the regression model are:
    a. Produce scatter plots of each independent variables against the dependent variable;
    b. Testing of the hypotheses on significance of the correlations between the dependent variable ticket sales and the independent variables;
    c. Test the significant regressions for multicolinearity, serial correlation and heteroscedacity;

d. Develop a multi-regression model based on the most important and significant independent variables and
e. Forecast the data for 12 months.

3.4 Evaluation of the models:

The forecasts of the appropriate models developed through the steps in paragraph 3.3 of the ticket sales of the period April 2007 – March 2008 are used for evaluation against the actual values recorded for the period.

The evaluation is conducted in the following steps:

1. Compare the different models of forecasting against the actual data for 2007/08;
2. Produce graphical representation of the various models and
3. Recommend the model with the lowest residuals or error terms that minimises the error of the forecasting model.

3.5 Combination of forecasts:

Based on the study of (Landram *et al.*,2008) and (Meade, 2000) two combination methods will be compared to establish whether any of these combination methods reduce the forecast error. The method of (Meade, 2000) entails using an equally weighted combination of forecasts and the method of (Landram *et al.*,2008) use the time series components of trend, seasonal and cyclical in combination with the best forecasting model.

The data variables and exploratory analysis of the data variable as well as the application of various quantitative models are dealt with in Chapter 4.

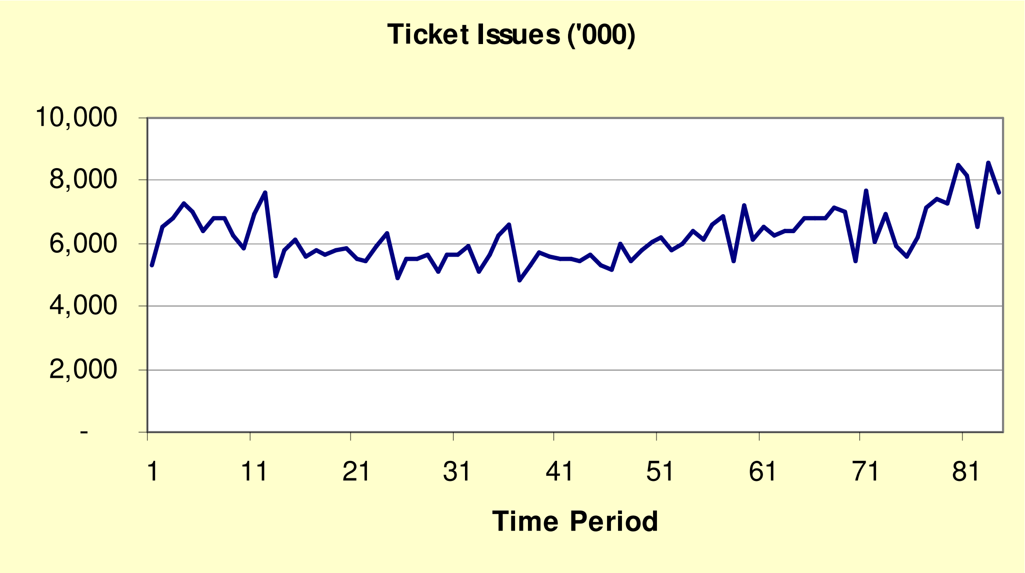
4 Study Results

The study results is discussed in four sections namely the

- Data variables and exploratory analysis of the data
- Quantitative models applied
- Comparison of quantitative models applied and
- Combination of forecasts.

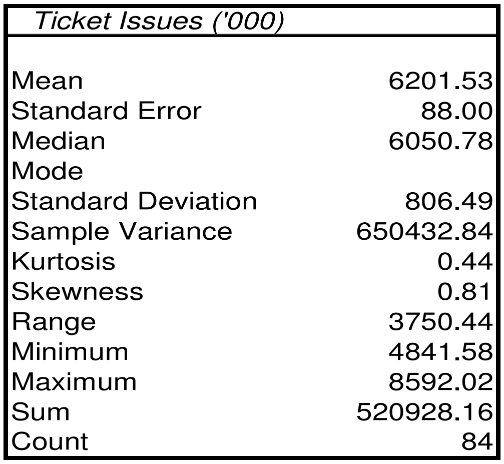
4.1 *Data variables and Exploratory Analysis of data used in the study:*

The ticket issues for the period of the study: April 2000 to March 2008, displays an upward trend as shown in the graph below:



Graph 1: Ticket Issues

The descriptive statistics for ticket issues are as follow:



Ticket Issues ('000)	
Mean	6201.53
Standard Error	88.00
Median	6050.78
Mode	
Standard Deviation	806.49
Sample Variance	650432.84
Kurtosis	0.44
Skewness	0.81
Range	3750.44
Minimum	4841.58
Maximum	8592.02
Sum	520928.16
Count	84

Table 1: Descriptive statistics: Ticket issues

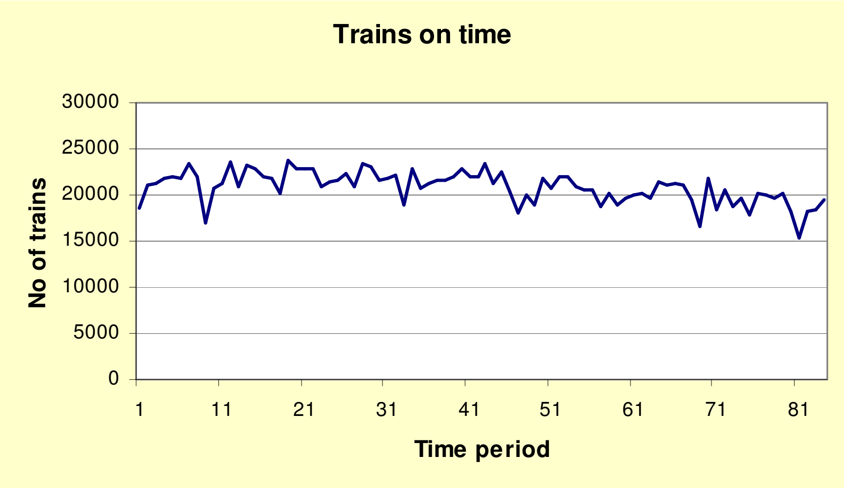
The data for 2007/08 was excluded in the application of the various models. The results of the various models were then compared with the data of 2007/08 to establish the most suitable forecasting method.

Explanatory variables gathered as independent variables are indicators of internal business factors and external factors impacting the business. The internal business indicators are trains on time, train cancellations, serious crime incidents and the price index of fare increases. Commuter surveys conducted by the organisation have indicated that safety and reliability are the main areas of importance for the commuter in terms of the commuter service (SARCC, 2006).

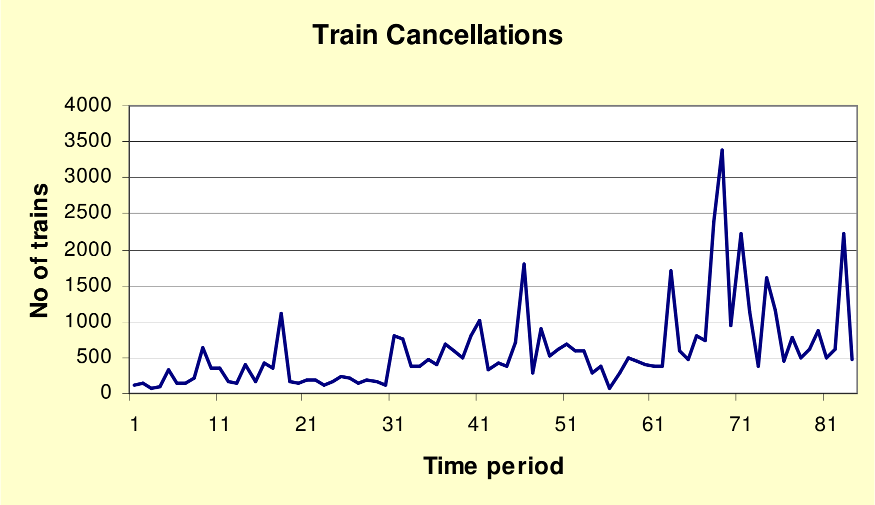
External indicators are the CPI (Inflation) and CPIX (Inflation excluding interest rates) as obtained from Statistics SA, and Fuel Prices for petrol, diesel and illuminating paraffin as obtained from the Department of Mineral and Energy Affairs. The fuel prices for coastal and inland regions have been averaged to provide a single indicator for the prices of petrol and diesel. Illuminating paraffin was selected as 45% of commuters are from households earning less than R2499 per month and illuminating paraffin is used in these households. (SARCC, 2006).

4.1.1 Internal indicators

Graphs of the internal indicators trains on time, trains cancelled, serious crime incidents and price index are provided to indicate the trends in the internal indicator variables.

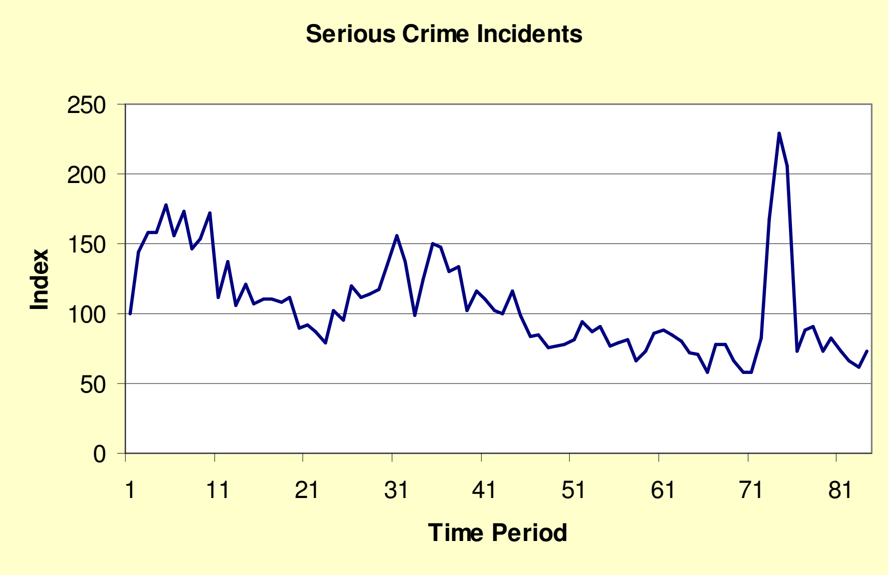


Graph 2: Trains on time

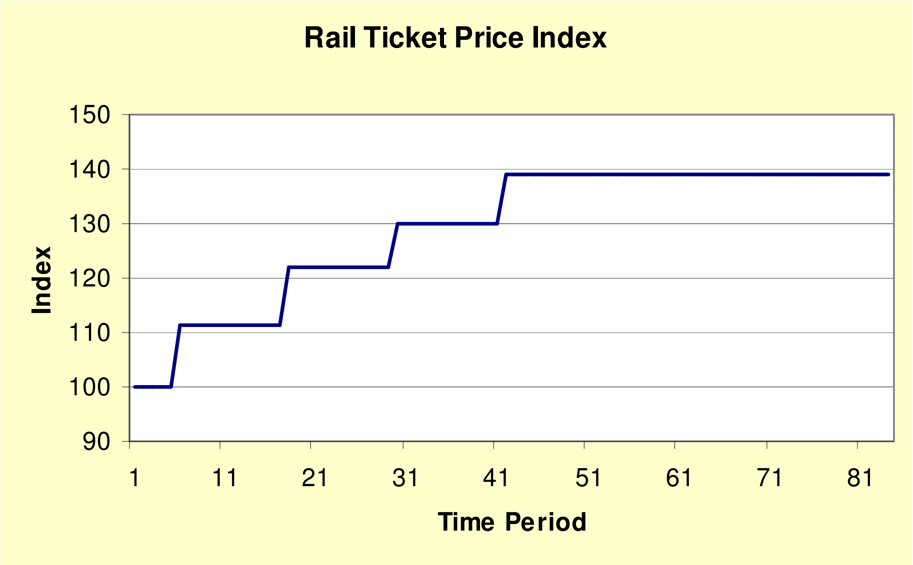


Graph 3: Trains Cancelled

.



Graph 4: Serious Crime Incidents

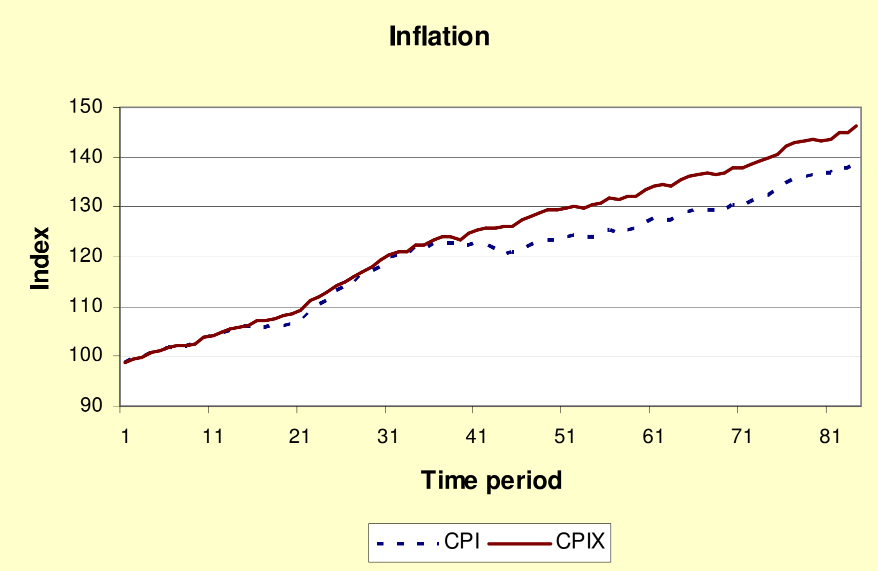


Graph 5: Ticket Price Index

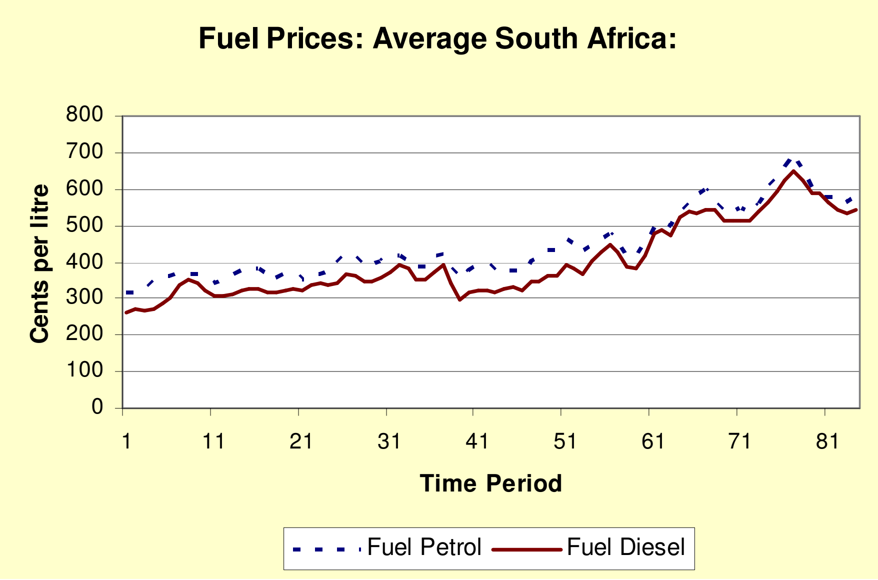
The next section will deal with the external variables selected for the regression analysis.

4.1.2 External variables:

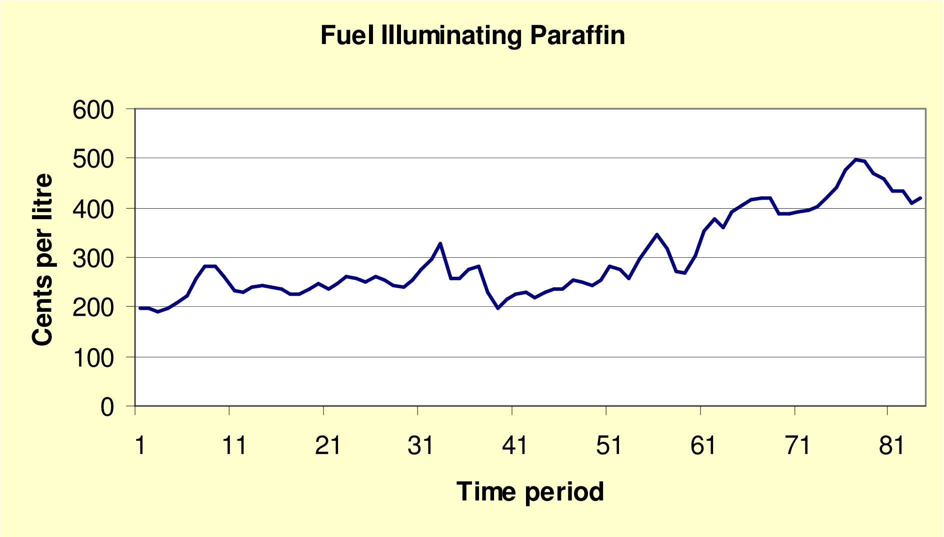
Graphical representations were done for the external explanatory variables.



Graph 6: Inflation

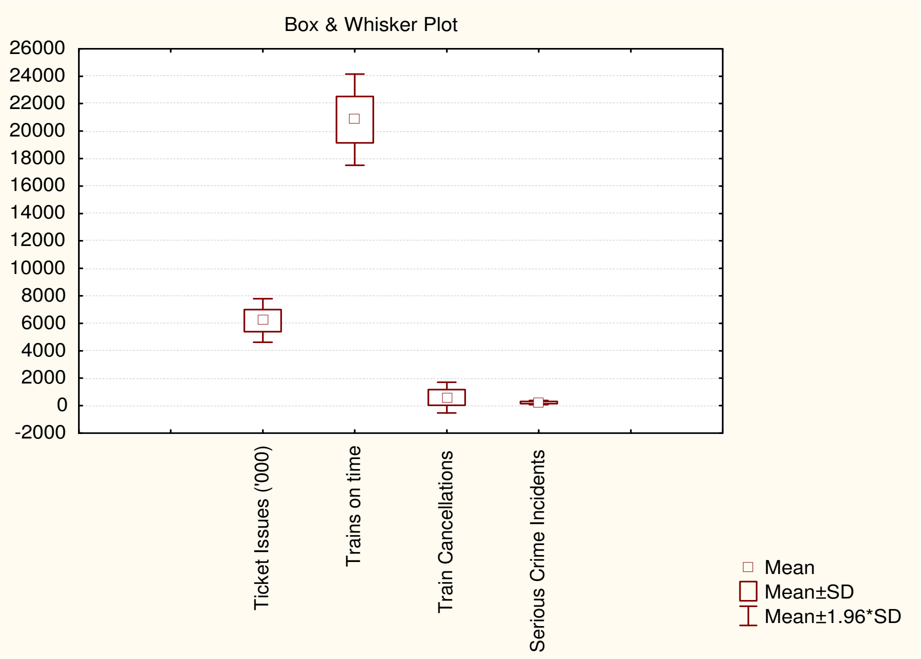


Graph 7: Average Fuel Prices: Petrol and Diesel

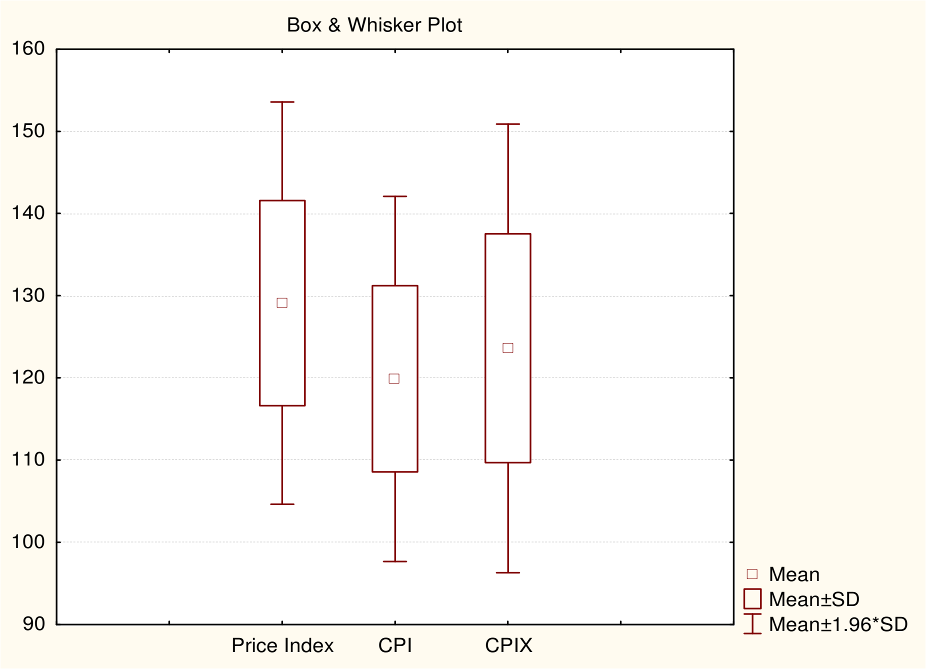


Graph 8: Prices: Illuminating Paraffin

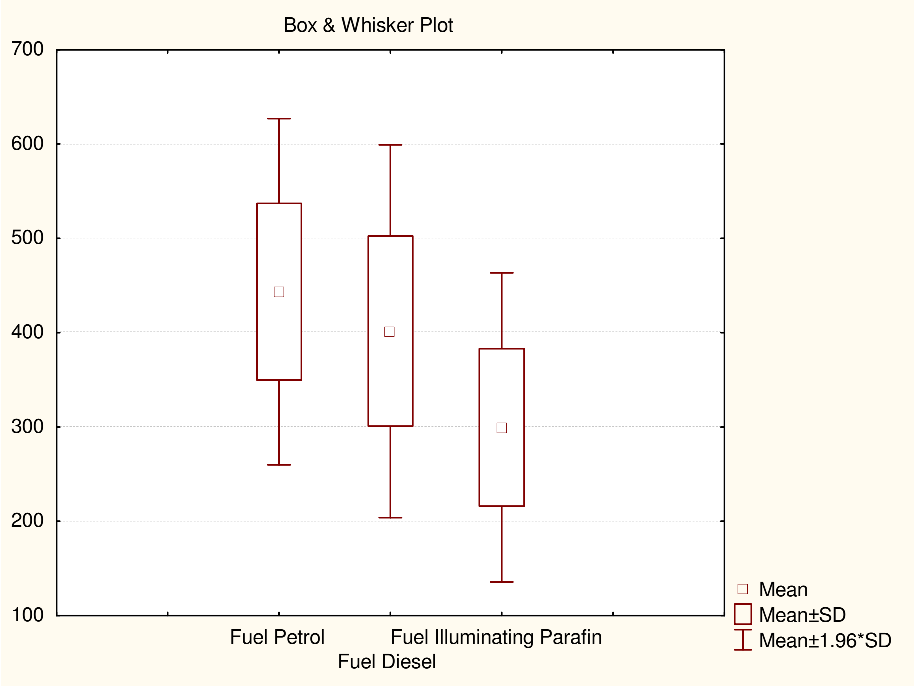
The variances as presented with Box and Whisker plot for Ticket issues, trains on time, train cancellations and serious crime incidents are presented in Graph 9. Variances for the index variables of Price Index, CPI (core inflation) and CPIX (inflation excluding interest rates) are presented in Graph 10. The variances for fuel price increases for the average petrol price, diesel price and illuminating paraffin price are presented in Graph 11.



Graph 9: Box and Whisker Plot: Ticket Issues, Trains on time, Train Cancellations and Serious Crime Incidents

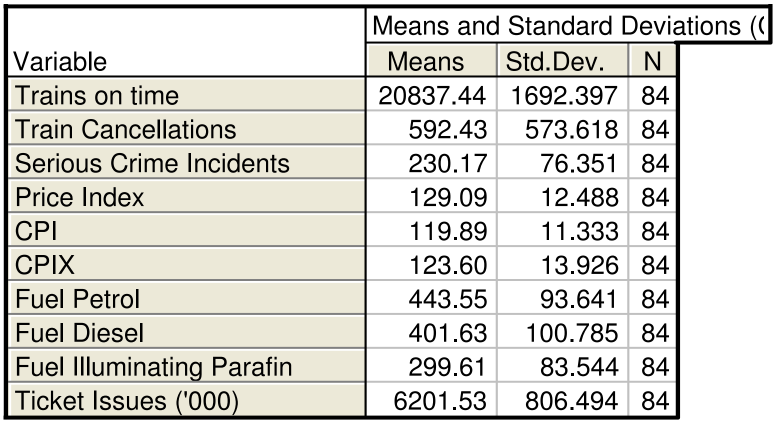


Graph 10: Box and Whisker Plot: Price Index, CPI and CPIX



Graph 11: Box and Whisker Plot: Average Fuel prices: Petrol, diesel and illuminating paraffin

The descriptive statistics of the variables are as follow:



	Means and Standard Deviations (		
Variable	Means	Std.Dev.	N
Trains on time	20837.44	1692.397	84
Train Cancellations	592.43	573.618	84
Serious Crime Incidents	230.17	76.351	84
Price Index	129.09	12.488	84
CPI	119.89	11.333	84
CPIX	123.60	13.926	84
Fuel Petrol	443.55	93.641	84
Fuel Diesel	401.63	100.785	84
Fuel Illuminating Parafin	299.61	83.544	84
Ticket Issues ('000)	6201.53	806.494	84

Table 2: Means and Standard Deviations for Regression variables

The values of the train cancellations presented large variances and was transformed with a log function prior to analysis.

The application of various quantitative statistical models is discussed in the next section.

4.2 Quantitative models applied.

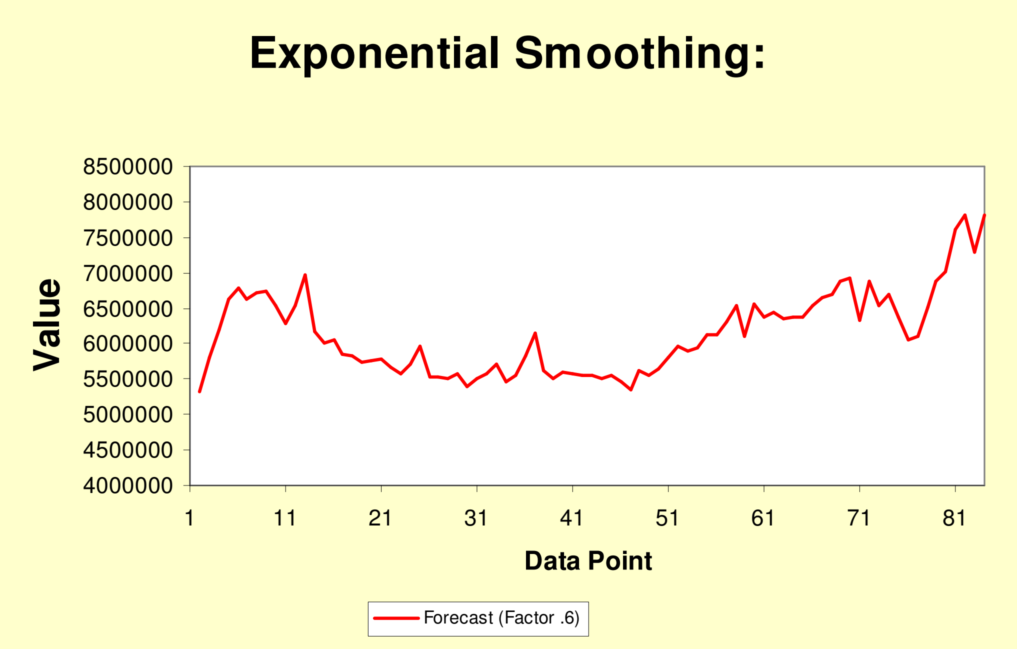
The results of the quantitative models applied are discussed in this section. The following models were addressed:

1. Time series decomposition and time series models.
2. The Box-Jenkins – ARIMA model
3. Regression and multi-regression modelling.

4.2.1 Time Series Decomposition for Commuter Rail data:

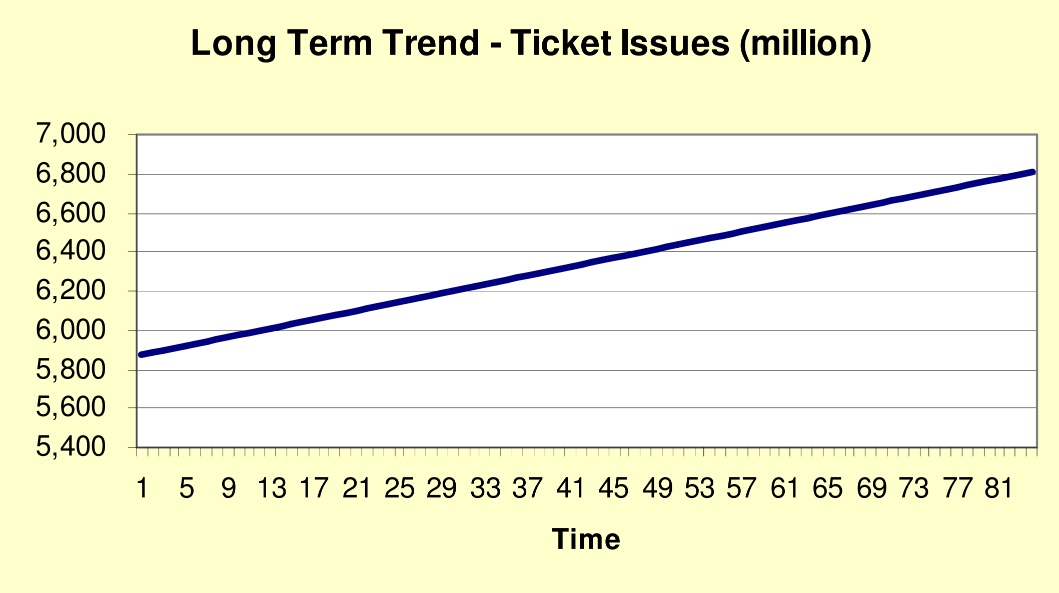
The data for ticket issues were decomposed as per decomposition analysis.

The trend of the data was obtained using exponential smoothing with a factor of 0.4. The exponential smoothing graph is as follow:



Graph 12: Exponential Smoothing: Ticket Issues

Determining the long term linear trend resulted in the following result:



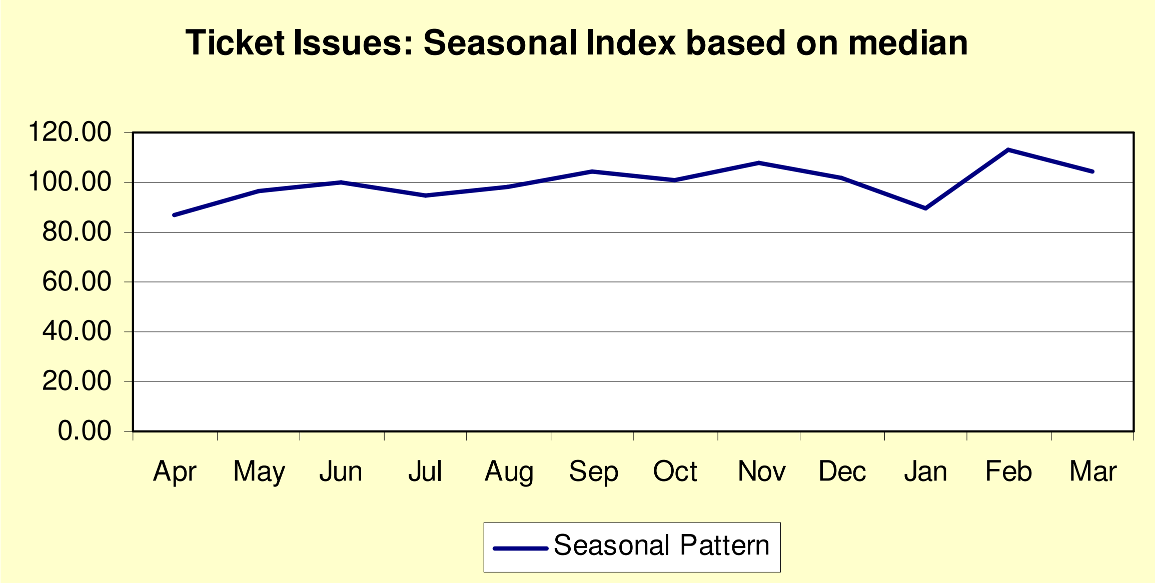
Graph 13: Long Term Linear Trend

The seasonal pattern was identified as follow:

	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	Seasonal median	Seasonally adjusted Index
Apr	5,319	4,987	4,893	4,842	5,773	6,516	6,927	5,319	87.18
May	6,519	5,788	5,517	5,329	6,062	6,225	5,910	5,910	96.87
Jun	6,782	6,102	5,494	5,716	6,174	6,385	5,591	6,102	100.01
Jul	7,305	5,555	5,639	5,575	5,801	6,375	6,168	5,801	95.08
Aug	7,014	5,774	5,109	5,531	6,007	6,789	7,142	6,007	98.46
Sep	6,412	5,617	5,668	5,513	6,391	6,792	7,391	6,391	104.76
Oct	6,830	5,804	5,672	5,466	6,150	6,770	7,247	6,150	100.80
Nov	6,791	5,822	5,905	5,613	6,580	7,169	8,504	6,580	107.85
Dec	6,230	5,484	5,112	5,325	6,864	6,976	8,144	6,230	102.11
Jan	5,880	5,444	5,667	5,185	5,443	5,446	6,507	5,446	89.26
Feb	6,920	5,911	6,273	6,008	7,223	7,706	8,592	6,920	113.42
Mar	7,605	6,358	6,613	5,468	6,142	6,040	7,623	6,358	104.20
								73,215	1200.00
							Average	6,101	100.00

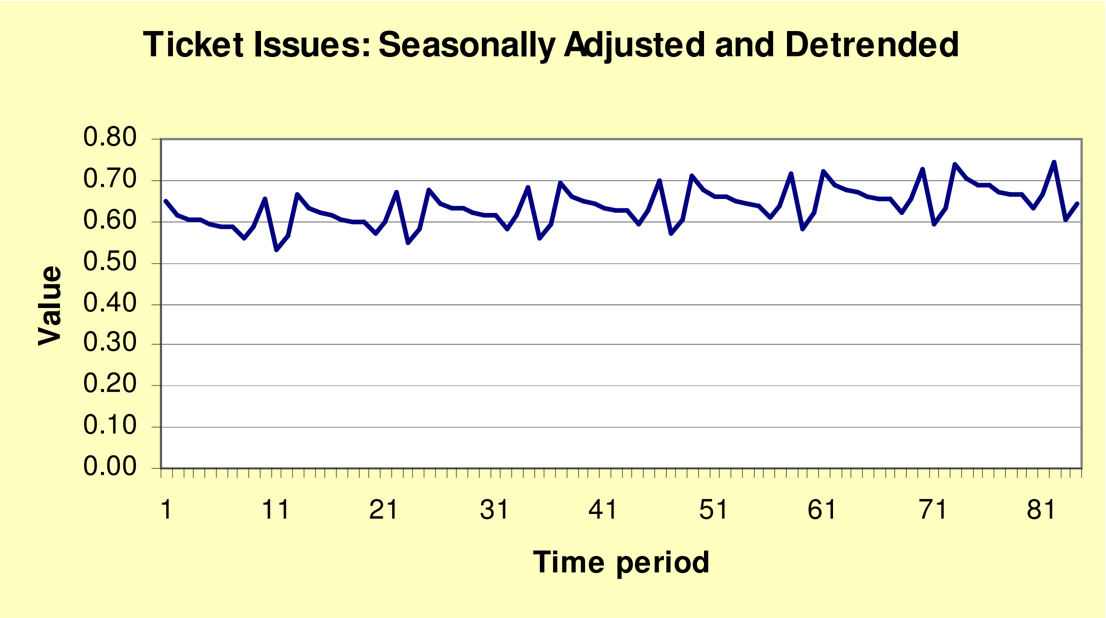
Table 3: Seasonal trend

The seasonal trend pattern graphically is as follow:



Graph 14: Seasonal Trend

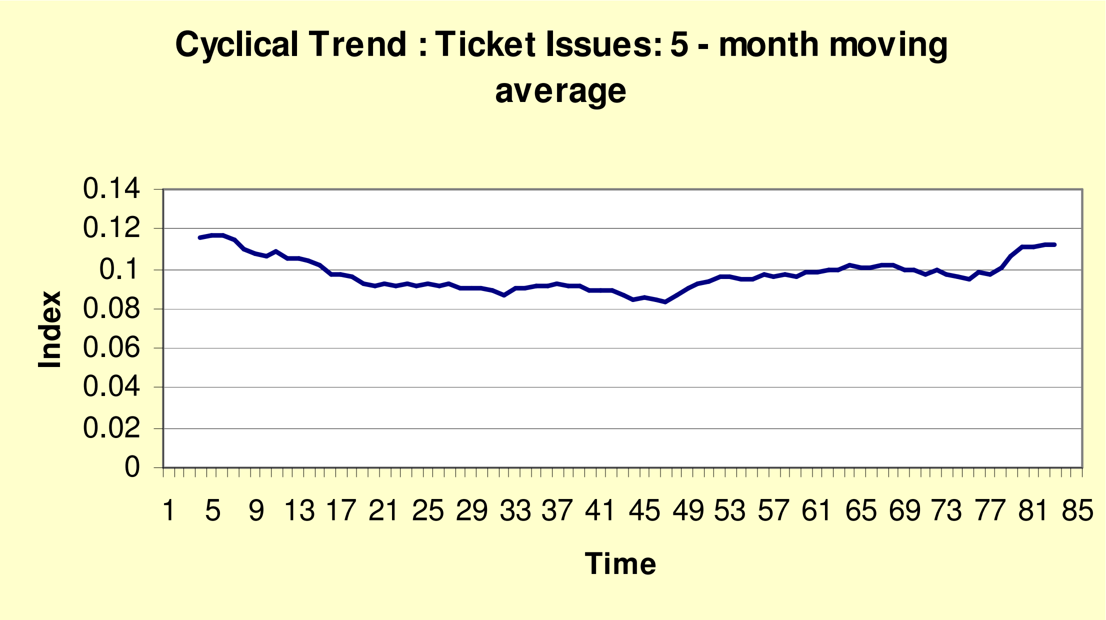
Adjusting the data with its seasonal pattern and eliminating the trend the data clearly highlighted the seasonal pattern as per Graph 15.



Graph 15: Seasonally adjusted and detrended data

Based on this pattern it is concluded that there is not only one season that differs from the others requiring the model for single-season heteroscadaticity.

The cyclical component was determined using the long term linear trend and is graphically depicted as follows:



Graph 16: Cyclical Trend

Graph 16 shows that there is not a real a cyclical pattern for the data. Graph 12 and 13 shows that there is an upward trend in the data series and from Graph 15 the deduction was made that there is a seasonal component to the data series.

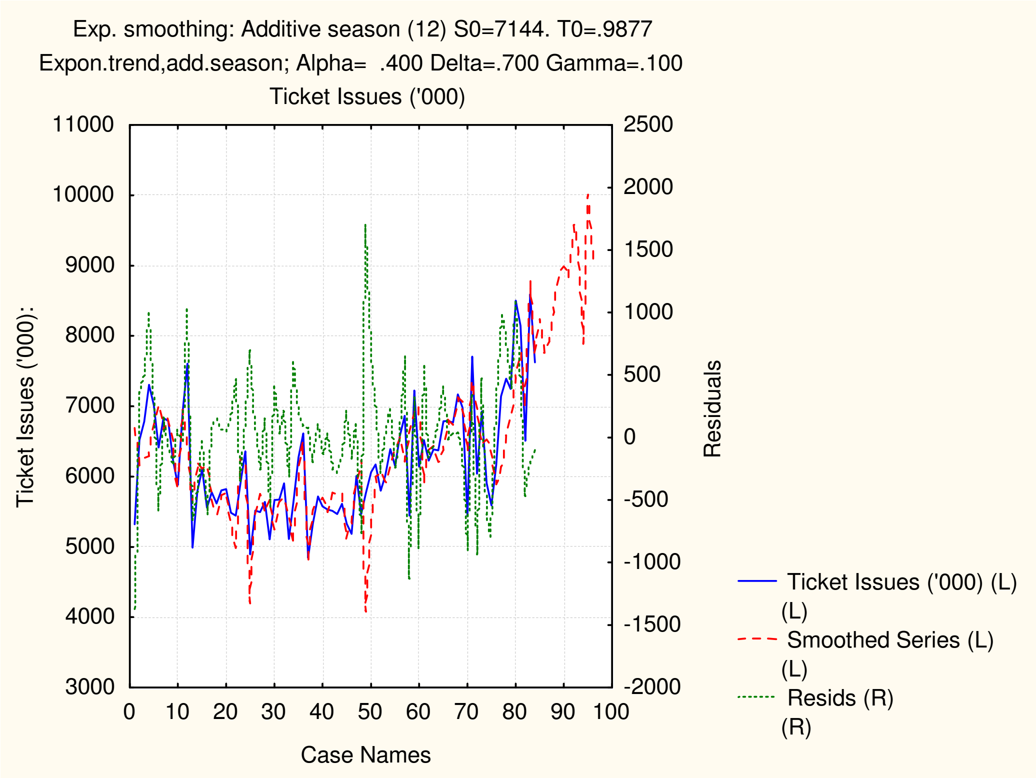
The following models were applied on the time series to determine the best fit model to the data series:

1. Exponential Smoothing additive model
2. Winter's Multiplicative model and
3. Exponential Smoothing multiplicative model.

The results of the time series models are discussed next.

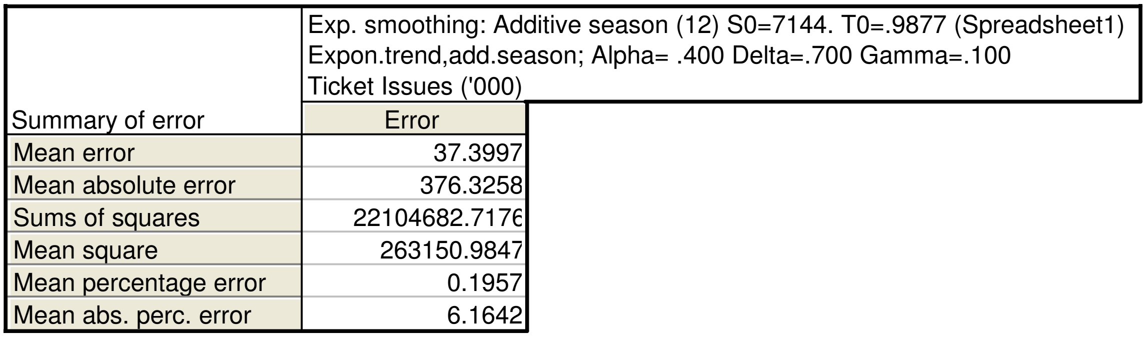
4.2.2 Exponential Smoothing Additive Model:

The plot and result of this model fitted to the data series ticket issues are as follow:



Graph 17: Exponential Smoothing Additive Model

The summary of errors is as follow:



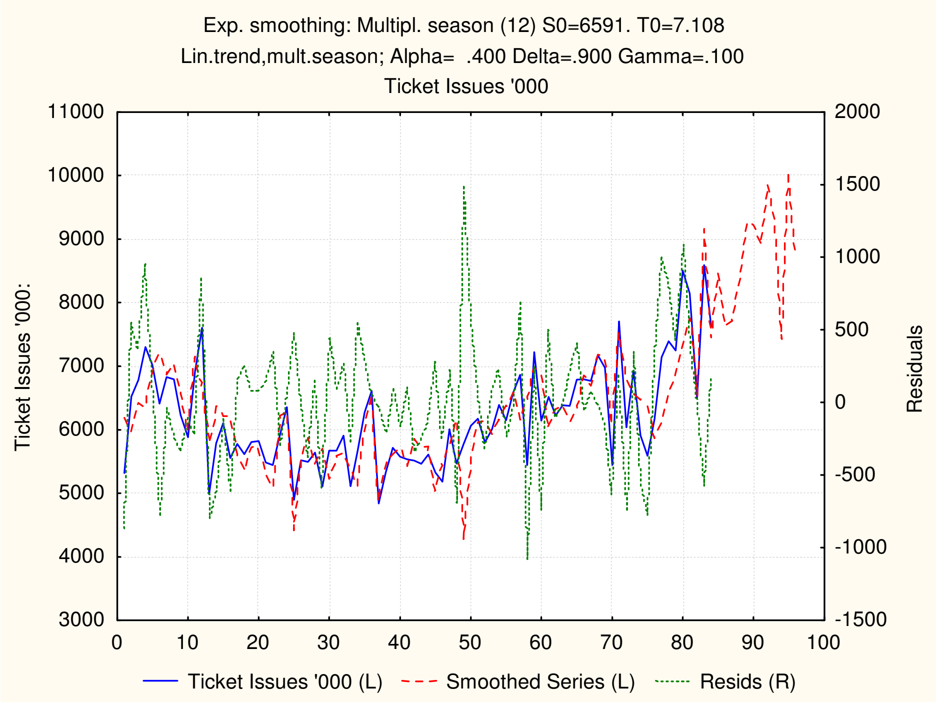
Summary of error	Exp. smoothing: Additive season (12) S0=7144. T0=.9877 (Spreadsheet1) Expon.trend,add.season; Alpha= .400 Delta=.700 Gamma=.100 Ticket Issues ('000) Error
Mean error	37.3997
Mean absolute error	376.3258
Sums of squares	22104682.7176
Mean square	263150.9847
Mean percentage error	0.1957
Mean abs. perc. error	6.1642

Table 4: Summary of error values: Exponential Smoothing Additive model

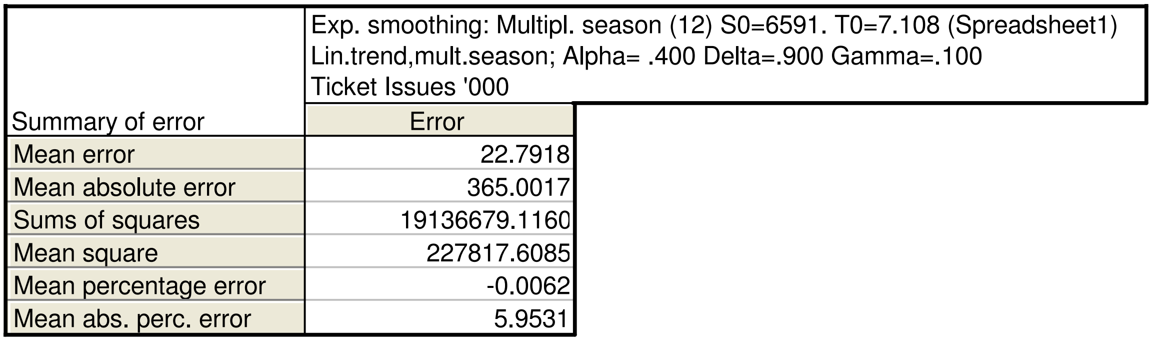
4.2.3 Winter's multiplicative model:

This model was applied as there is a seasonal trend in the data. The parameters indicating the lowest errors for three error measures obtained using a grid search was applied to the model.

The plot and error values are as follow:



Graph 18: Winter's Multiplicative Model

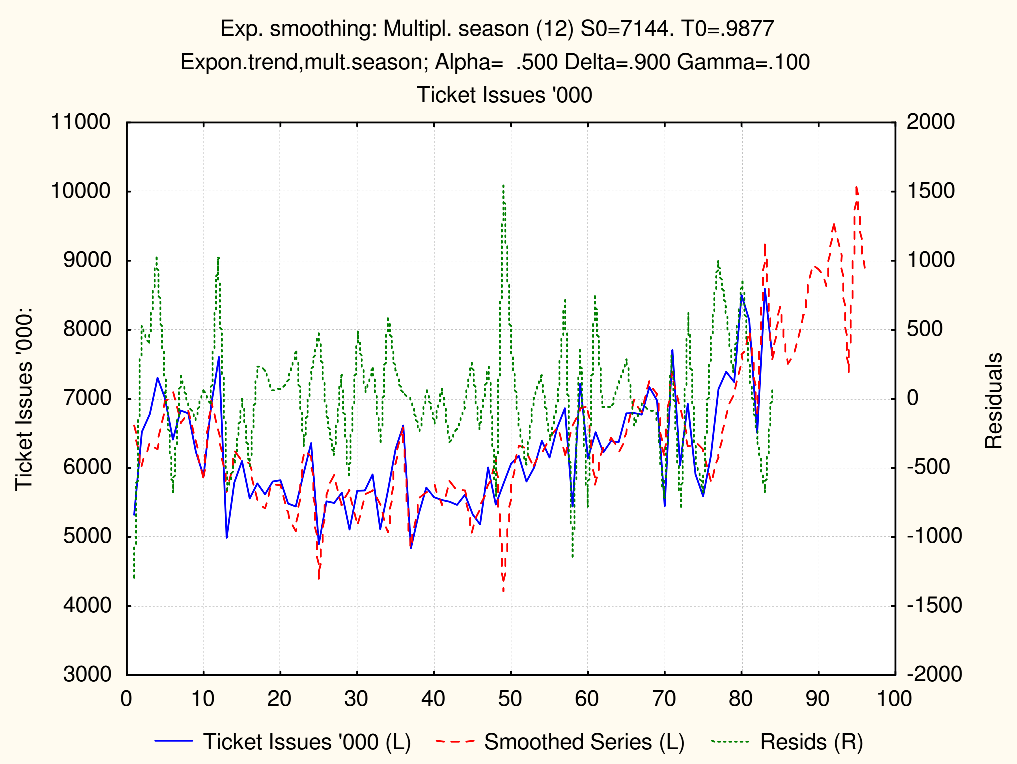


Summary of error	Exp. smoothing: Multipl. season (12) S0=6591. T0=7.108 (Spreadsheet1) Lin.trend,mult.season; Alpha= .400 Delta=.900 Gamma=.100 Ticket Issues '000 Error
Mean error	22.7918
Mean absolute error	365.0017
Sums of squares	19136679.1160
Mean square	227817.6085
Mean percentage error	-0.0062
Mean abs. perc. error	5.9531

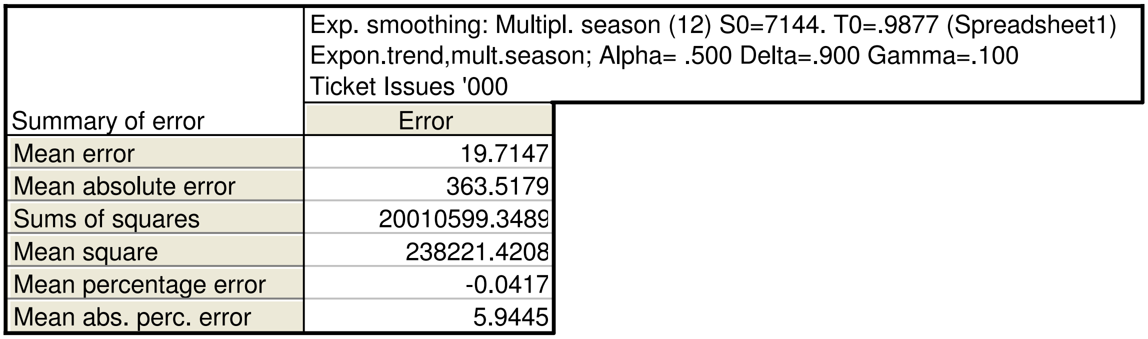
Table 5: Winter's Multiplicative Model: Error values

4.2.4 Exponential Smoothing Multiplicative Model:

A grid search was used to determine the smallest mean error for the exponential smoothing multiplicative model. The parameters for this smallest mean error were applied to the model. The plot and error values are as follow:



Graph 19: Exponential Smoothing Multiplicative Model



	Exp. smoothing: Multipl. season (12) S0=7144. T0=.9877 (Spreadsheet1) Expon.trend,mult.season; Alpha= .500 Delta=.900 Gamma=.100 Ticket Issues '000
Summary of error	Error
Mean error	19.7147
Mean absolute error	363.5179
Sums of squares	20010599.3489
Mean square	238221.4208
Mean percentage error	-0.0417
Mean abs. perc. error	5.9445

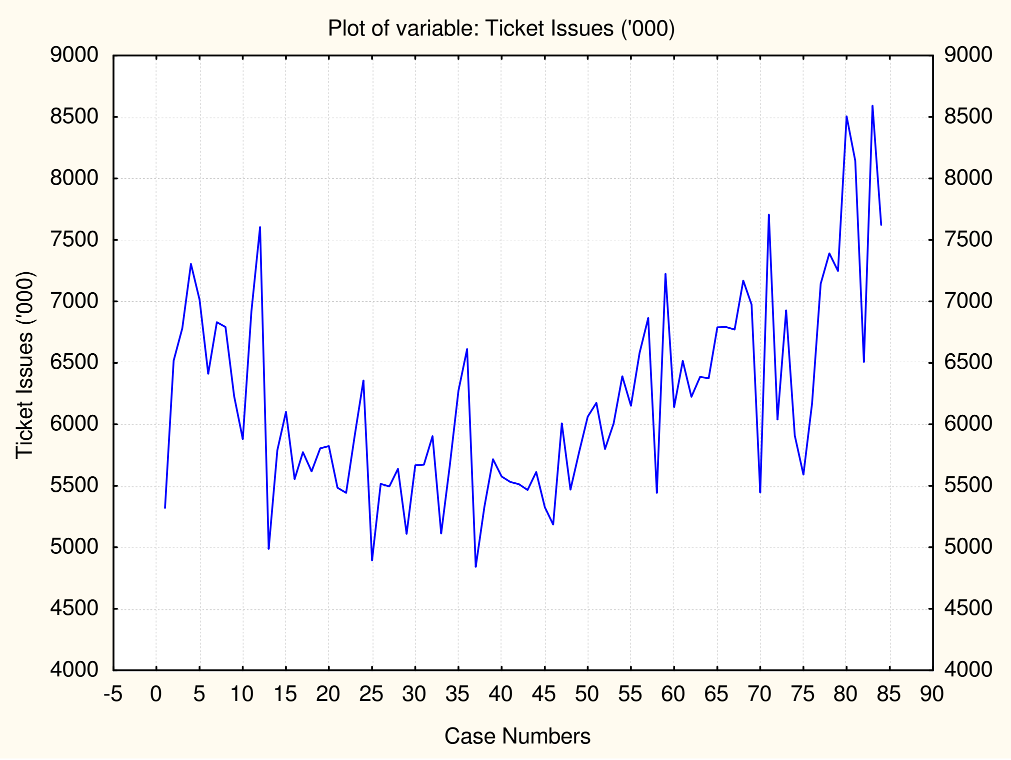
Table 6: Exponential Smoothing Multiplicative Model: Error values

4.2.5 Comparison of the time series models:

The three models all produced low mean error values. The model with the lowest mean error is the Exponential Smoothing Multiplicative model and should therefore present the best results. All three models will however be analysed for accuracy in forecasting for the ticket issues for April 2007 – March 2008.

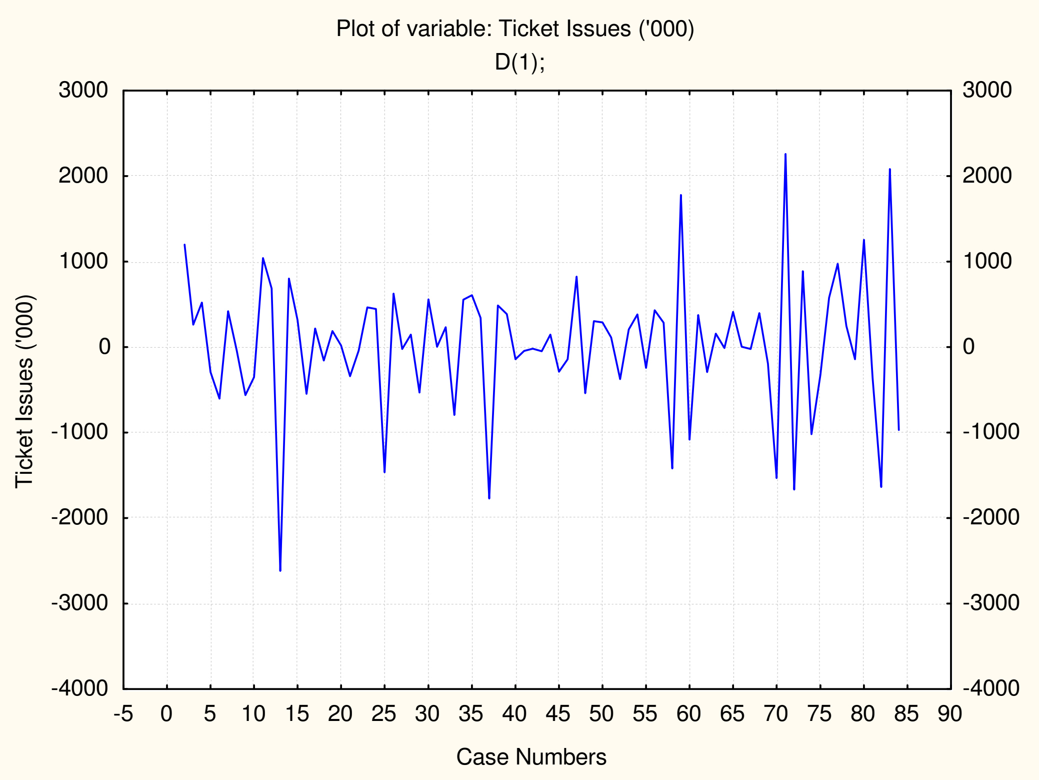
4.2.6 Box-Jenkins or ARIMA Model:

The first step for the ARIMA model is to ensure that the series to be fitted are stationary. The graph below shows the original data before differentiation and it is clear from the graph that the series is not stationary.



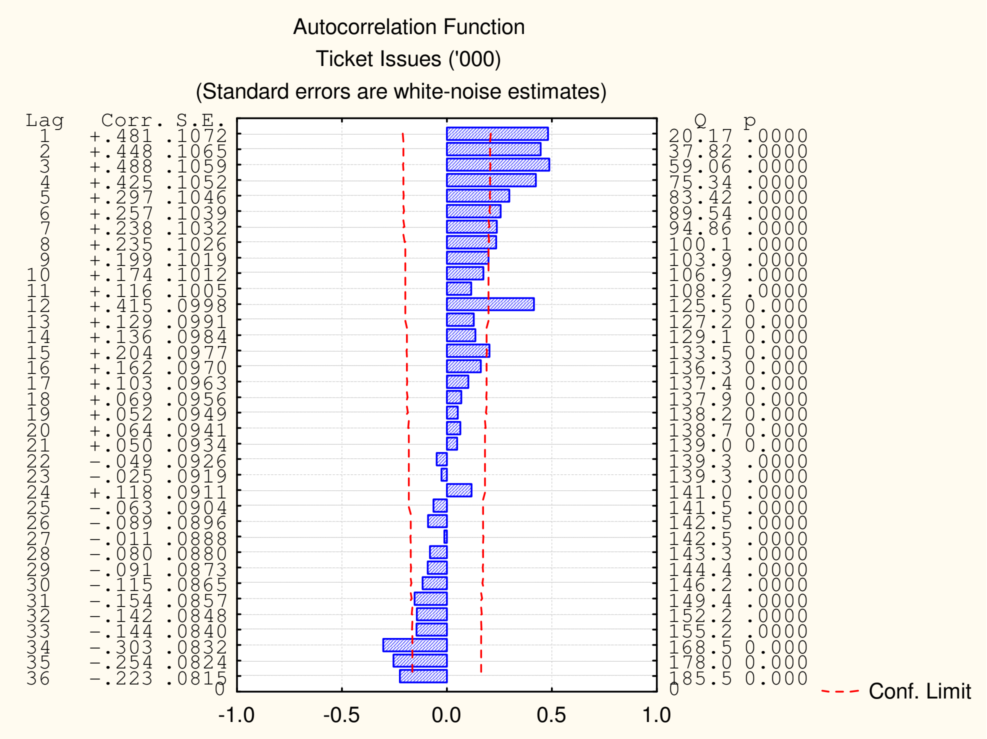
Graph 20: Ticket Issues prior to differentiation

The data was then differentiated with one lag to make the series stationary as per Graph 21.



Graph 21: Differenced Ticket Issues - Lag 1

The autocorrelation of the variable Ticket issues prior to differentiation is as presented in Graph 22.

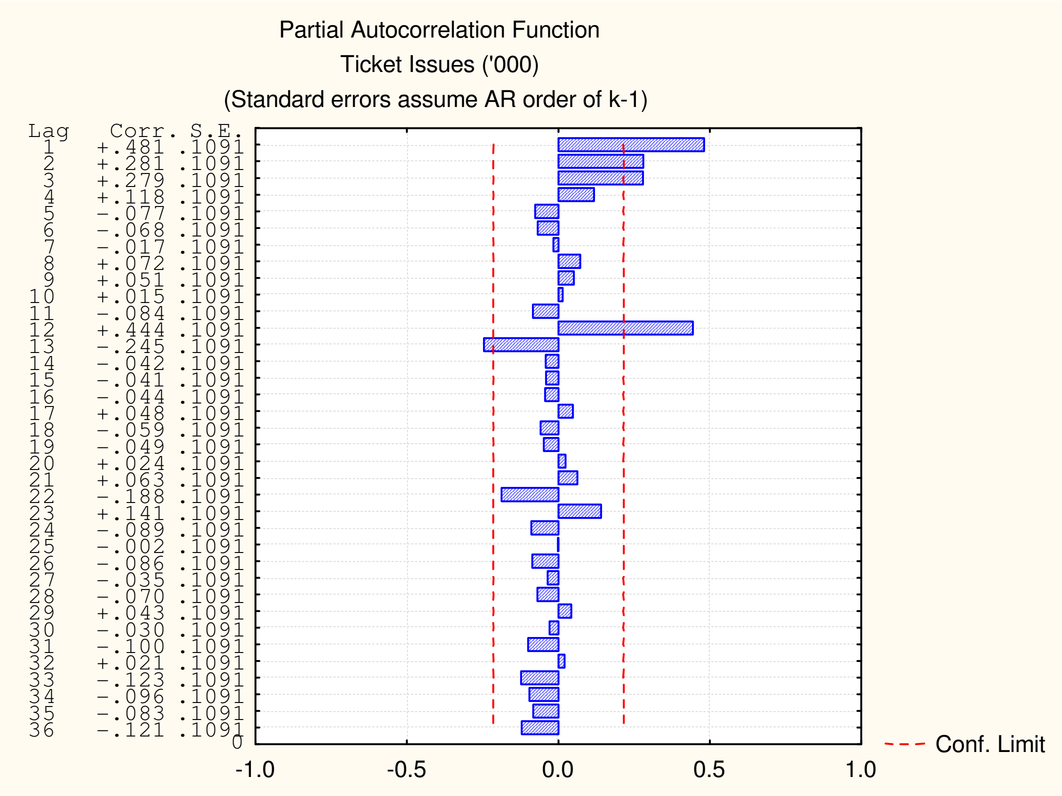


Lag	Corr.	S.E.	Q	p
1	+.481	.1072	20.17	.0000
2	+.448	.1065	37.82	.0000
3	+.488	.1059	59.06	.0000
4	+.425	.1052	75.34	.0000
5	+.297	.1046	83.42	.0000
6	+.257	.1039	89.54	.0000
7	+.238	.1032	94.86	.0000
8	+.235	.1026	100.1	.0000
9	+.199	.1019	103.9	.0000
10	+.174	.1012	106.9	.0000
11	+.116	.1005	108.2	.0000
12	+.415	.0998	125.5	0.000
13	+.129	.0991	127.2	0.000
14	+.136	.0984	129.1	0.000
15	+.204	.0977	133.5	0.000
16	+.162	.0970	136.3	0.000
17	+.103	.0963	137.4	0.000
18	+.069	.0956	137.9	0.000
19	+.052	.0949	138.2	0.000
20	+.064	.0941	138.7	0.000
21	+.050	.0934	139.0	0.000
22	-.049	.0926	139.3	.0000
23	-.025	.0919	139.3	.0000
24	+.118	.0911	141.0	.0000
25	-.063	.0904	141.5	.0000
26	-.089	.0896	142.5	.0000
27	-.011	.0888	142.5	.0000
28	-.080	.0880	143.3	.0000
29	-.091	.0873	144.4	.0000
30	-.115	.0865	146.2	.0000
31	-.154	.0857	149.4	.0000
32	-.142	.0848	152.2	.0000
33	-.144	.0840	155.2	.0000
34	-.303	.0832	168.5	0.000
35	-.254	.0824	178.0	0.000
36	-.223	.0815	185.5	0.000

Graph 22: Autocorrelation: Ticket Issues

The autocorrelation function indicates a seasonal pattern for the data and in addition the autocorrelations seem to follow a sine-wave pattern.

The partial autocorrelation for the series prior to differentiation is presented in Graph 23. The partial autocorrelation has spikes at lags 1,2 and 3 and then again significant partial correlations at lags 11 and 12 that are indicative of the seasonal pattern.



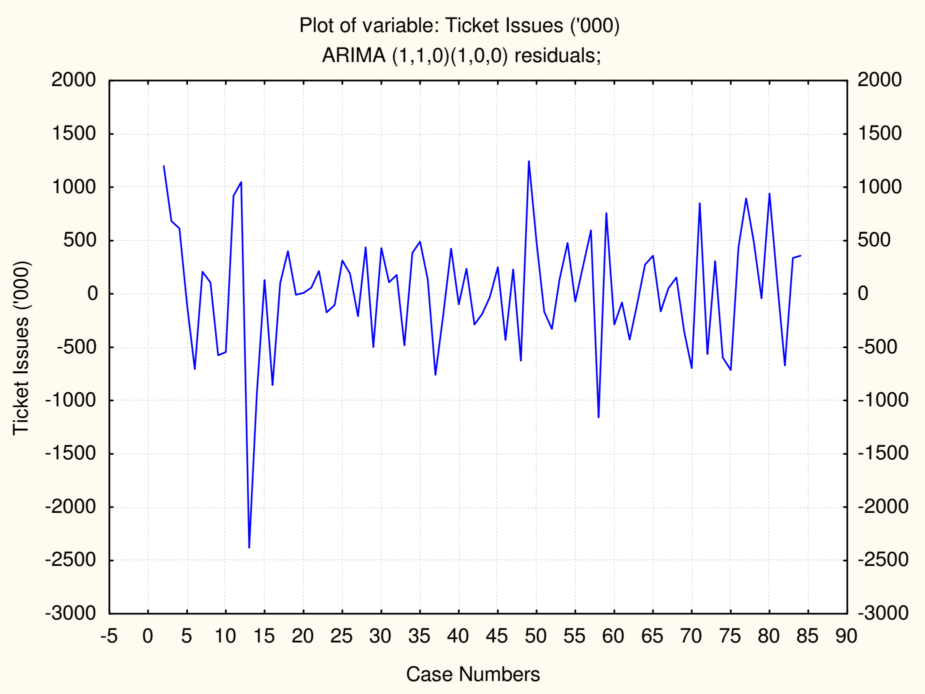
Graph 23: Partial Autocorrelation: Ticket Issues

Based on the pattern of the autocorrelations and partial autocorrelation, an ARIMA (1,0,0)(1,0,0) model was applied to the data. The results at a 95% confidence level and 5% significance level are as follow:

	Input: Ticket Issues ('000) (Ticket issues) Transformations: D(1) Model:(1,1,0)(1,0,0) Seasonal lag: 12 MS Residual= 3282E2					
Paramet.	Param.	Asympt. Std.Err.	Asympt. t(81)	p	Lower 95% Conf	Upper 95% Conf
p(1)	-0.3486	0.1083	-3.2179	0.0019	-0.5641	-0.1330
Ps(1)	0.6819	0.1067	6.3918	0.0000	0.4697	0.8942

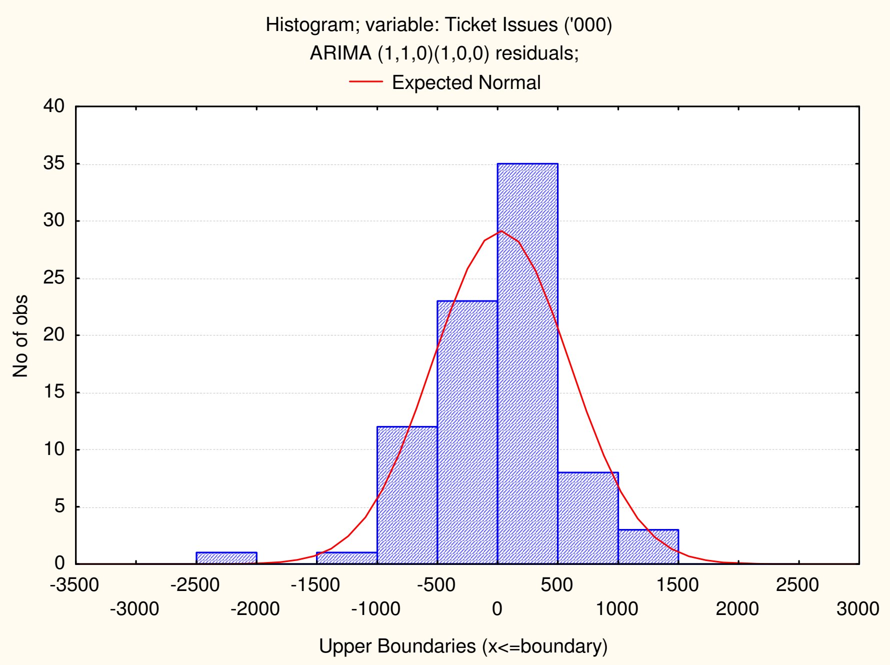
Table 7: Parameters for ARIMA (1,1,0)(1,0,0) model

The residuals of the model are presented in Graph 24.



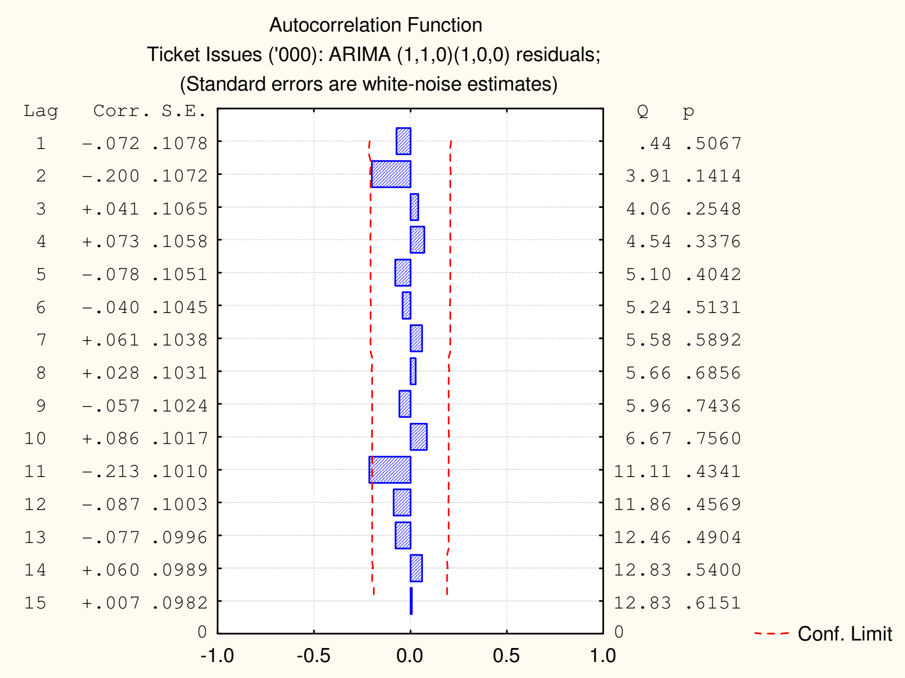
Graph 24: Residuals of ARIMA (1,1,0)(1,0,0)

The distribution of the residuals approximates the normal distribution although with some residuals outside the expected normal distribution.

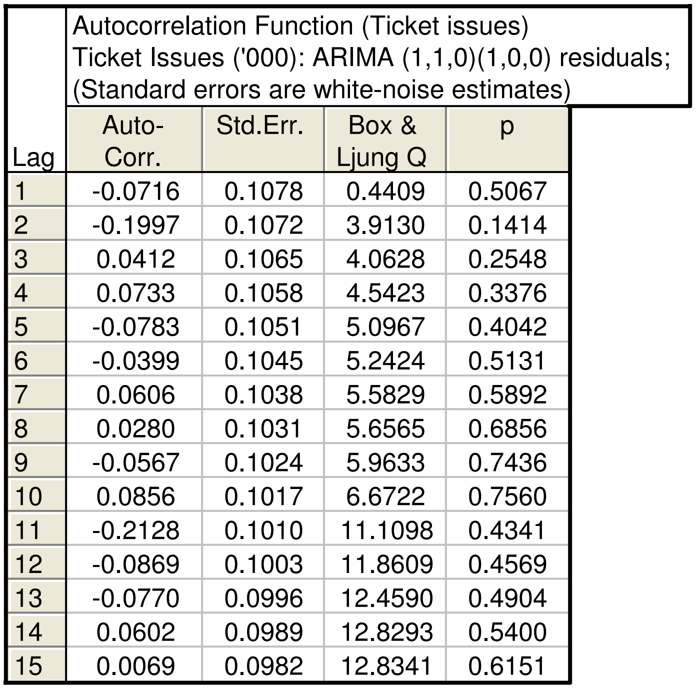


Graph 25: Distribution of ARIMA (1,1,0)(1,0,0) residuals.

In addition there was no significant auto-correlation found for the residuals or a discernable serial pattern. A significant partial autocorrelation was found at lag 11. Therefore the model might not be adequate for the forecast.



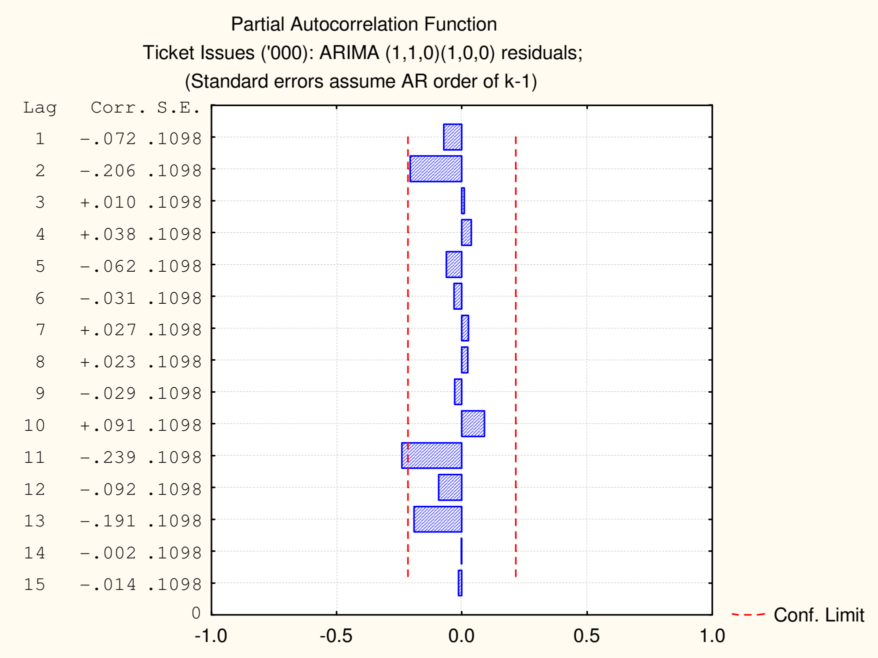
Graph 26: Autocorrelation: ARIMA (1,1,0)(1,0,0) Residuals



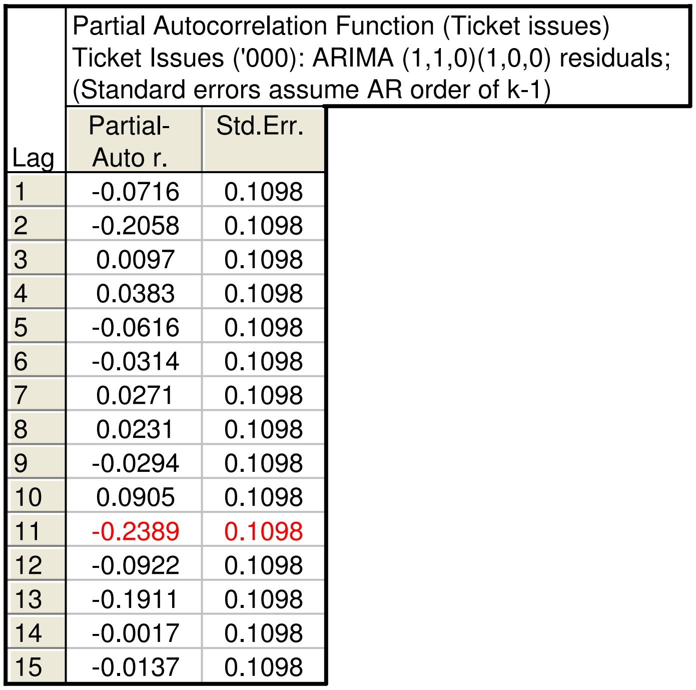
Autocorrelation Function (Ticket issues)
Ticket Issues ('000): ARIMA (1,1,0)(1,0,0) residuals;
(Standard errors are white-noise estimates)

Lag	Auto-Corr.	Std.Err.	Box & Ljung Q	p
1	-0.0716	0.1078	0.4409	0.5067
2	-0.1997	0.1072	3.9130	0.1414
3	0.0412	0.1065	4.0628	0.2548
4	0.0733	0.1058	4.5423	0.3376
5	-0.0783	0.1051	5.0967	0.4042
6	-0.0399	0.1045	5.2424	0.5131
7	0.0606	0.1038	5.5829	0.5892
8	0.0280	0.1031	5.6565	0.6856
9	-0.0567	0.1024	5.9633	0.7436
10	0.0856	0.1017	6.6722	0.7560
11	-0.2128	0.1010	11.1098	0.4341
12	-0.0869	0.1003	11.8609	0.4569
13	-0.0770	0.0996	12.4590	0.4904
14	0.0602	0.0989	12.8293	0.5400
15	0.0069	0.0982	12.8341	0.6151

Table 8: Autocorrelation for residuals: ARIMA(1,1,0)(1,0,0)



Graph 27: Partial Autocorrelation: ARIMA(1,1,0)(1,0,0) residuals

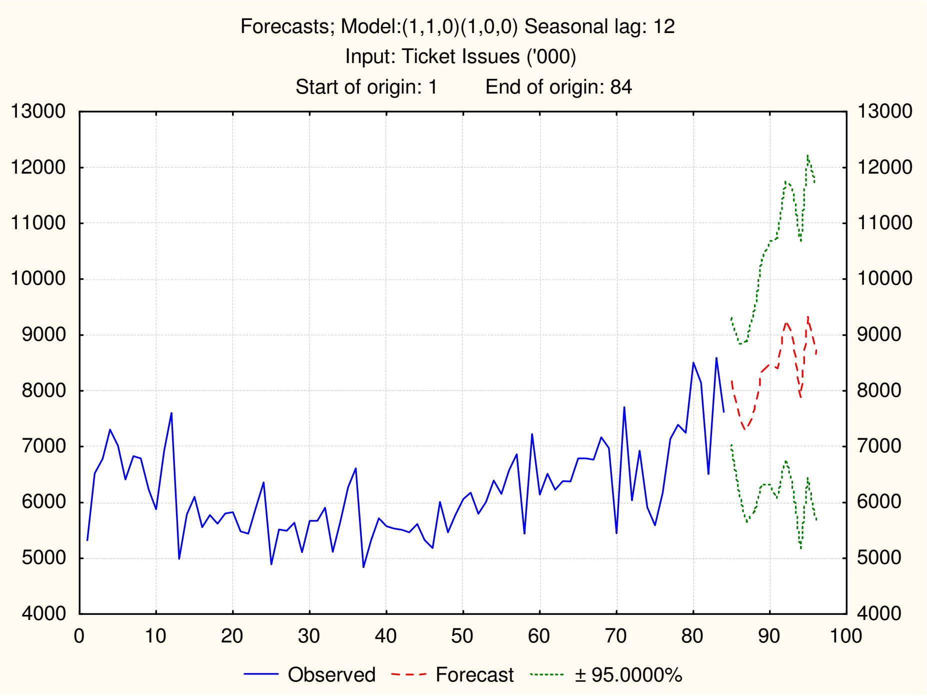


Partial Autocorrelation Function (Ticket issues)
Ticket Issues ('000): ARIMA (1,1,0)(1,0,0) residuals;
(Standard errors assume AR order of k-1)

Lag	Partial-Auto r.	Std.Err.
1	-0.0716	0.1098
2	-0.2058	0.1098
3	0.0097	0.1098
4	0.0383	0.1098
5	-0.0616	0.1098
6	-0.0314	0.1098
7	0.0271	0.1098
8	0.0231	0.1098
9	-0.0294	0.1098
10	0.0905	0.1098
11	-0.2389	0.1098
12	-0.0922	0.1098
13	-0.1911	0.1098
14	-0.0017	0.1098
15	-0.0137	0.1098

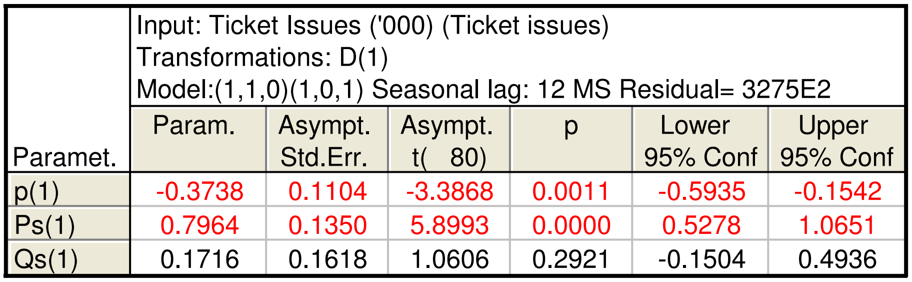
Table 9: Partial Autocorrelation Function Residuals: ARIMA(1,1,0)(1,0,0)

The forecasts of the model are as follow:



Graph 28: Forecasts of ARIMA(1,1,0)(1,0,0)

As this model still showed a significant partial autocorrelation, a seasonal moving average parameter was added to the model. An ARIMA (1,1,0)(1,0,1) was applied with the following parameters:

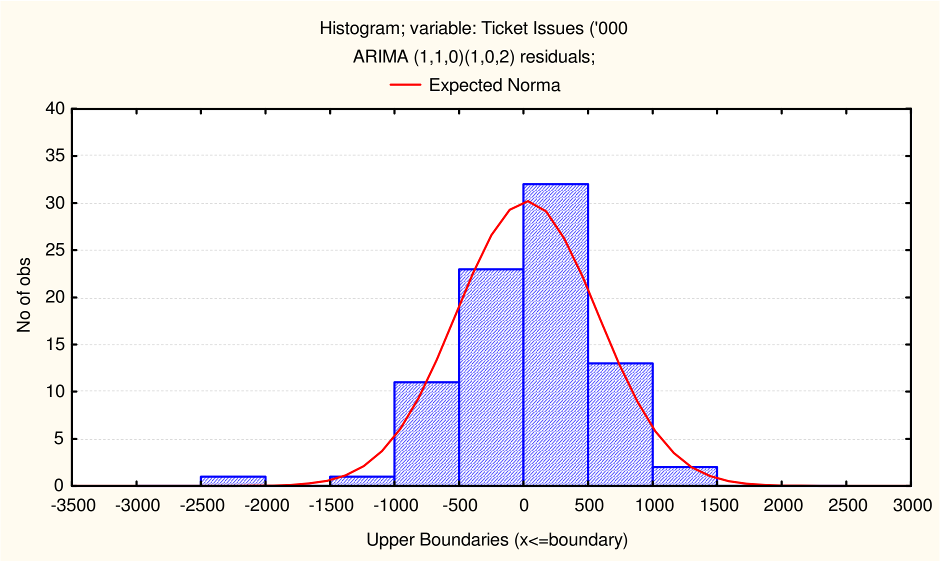


	Input: Ticket Issues ('000) (Ticket issues) Transformations: D(1) Model:(1,1,0)(1,0,1) Seasonal lag: 12 MS Residual= 3275E2					
Paramet.	Param.	Asympt. Std.Err.	Asympt. t(80)	p	Lower 95% Conf	Upper 95% Conf
p(1)	-0.3738	0.1104	-3.3868	0.0011	-0.5935	-0.1542
Ps(1)	0.7964	0.1350	5.8993	0.0000	0.5278	1.0651
Qs(1)	0.1716	0.1618	1.0606	0.2921	-0.1504	0.4936

Table 10: Results for ARIMA (1,1,0)(1,0,1)

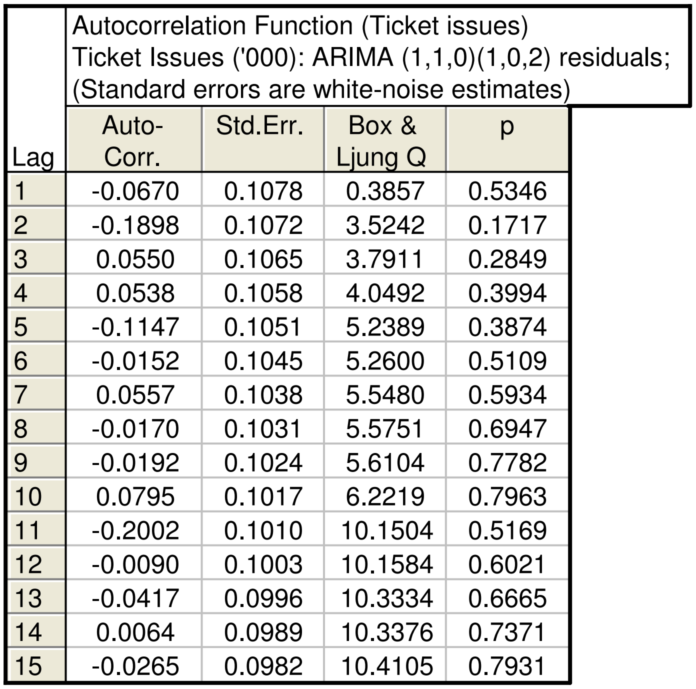
The t-value for the moving average seasonal parameter is not significant. Therefore the moving average seasonal parameter was increased to 2. The result of this adjustment provided a better result on the autocorrelation function and the partial autocorrelation function.

The distribution of the residuals of the ARIMA(1,1,0)(1,0,2) model is as follow:



Graph 29: Distribution of the residuals: ARIMA(1,1,0)(1,0,2) model

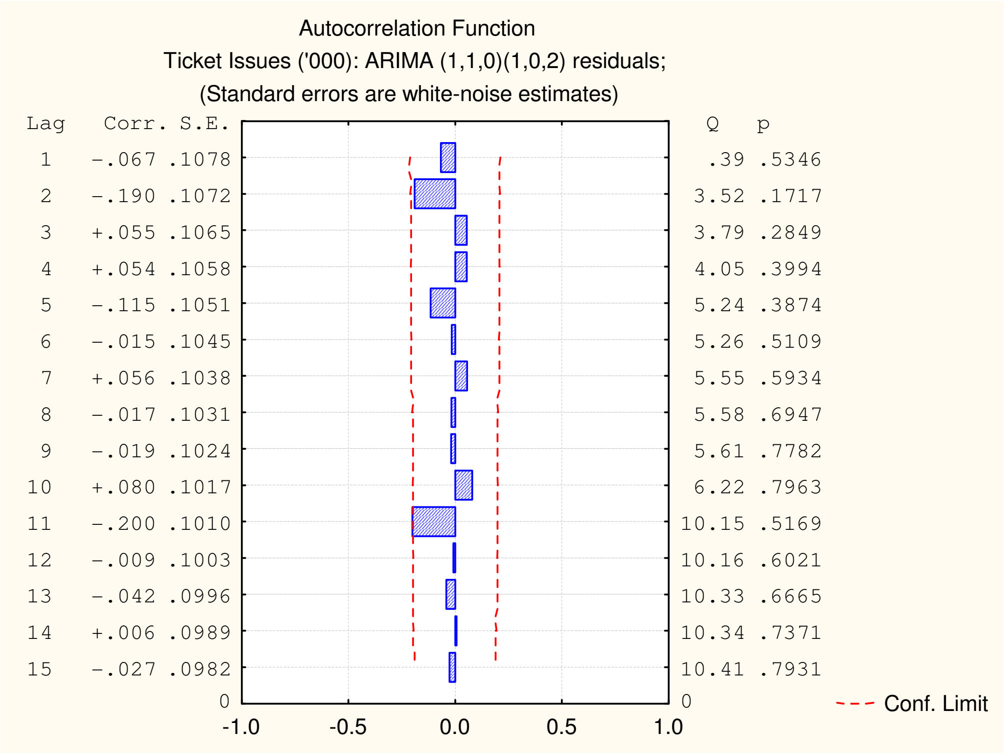
The autocorrelation function for the residuals has no significant correlations:



Autocorrelation Function (Ticket issues)
Ticket Issues ('000): ARIMA (1,1,0)(1,0,2) residuals;
(Standard errors are white-noise estimates)

Lag	Auto- Corr.	Std.Err.	Box & Ljung Q	p
1	-0.0670	0.1078	0.3857	0.5346
2	-0.1898	0.1072	3.5242	0.1717
3	0.0550	0.1065	3.7911	0.2849
4	0.0538	0.1058	4.0492	0.3994
5	-0.1147	0.1051	5.2389	0.3874
6	-0.0152	0.1045	5.2600	0.5109
7	0.0557	0.1038	5.5480	0.5934
8	-0.0170	0.1031	5.5751	0.6947
9	-0.0192	0.1024	5.6104	0.7782
10	0.0795	0.1017	6.2219	0.7963
11	-0.2002	0.1010	10.1504	0.5169
12	-0.0090	0.1003	10.1584	0.6021
13	-0.0417	0.0996	10.3334	0.6665
14	0.0064	0.0989	10.3376	0.7371
15	-0.0265	0.0982	10.4105	0.7931

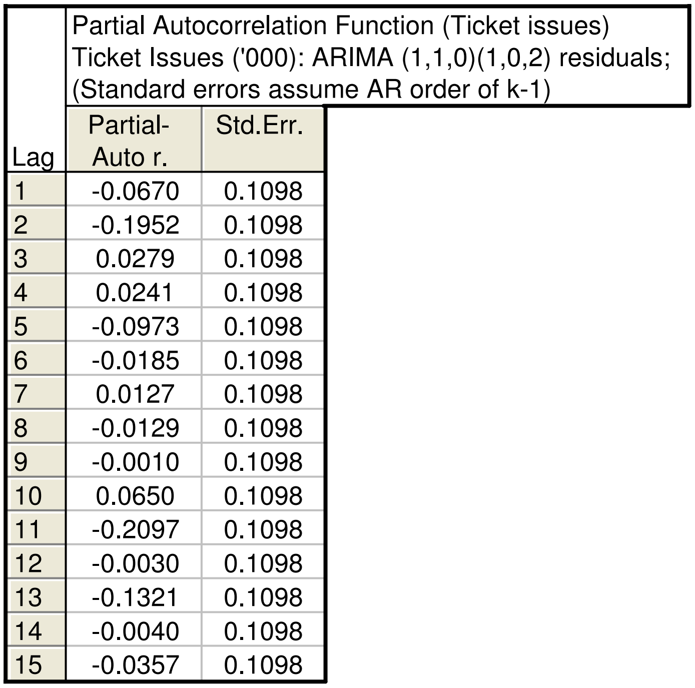
Table 11: Autocorrelations of Residuals: ARIMA(1,1,0)(1,0,2)



Lag	Corr.	S.E.	Q	p
1	-.067	.1078	.39	.5346
2	-.190	.1072	3.52	.1717
3	+.055	.1065	3.79	.2849
4	+.054	.1058	4.05	.3994
5	-.115	.1051	5.24	.3874
6	-.015	.1045	5.26	.5109
7	+.056	.1038	5.55	.5934
8	-.017	.1031	5.58	.6947
9	-.019	.1024	5.61	.7782
10	+.080	.1017	6.22	.7963
11	-.200	.1010	10.15	.5169
12	-.009	.1003	10.16	.6021
13	-.042	.0996	10.33	.6665
14	+.006	.0989	10.34	.7371
15	-.027	.0982	10.41	.7931

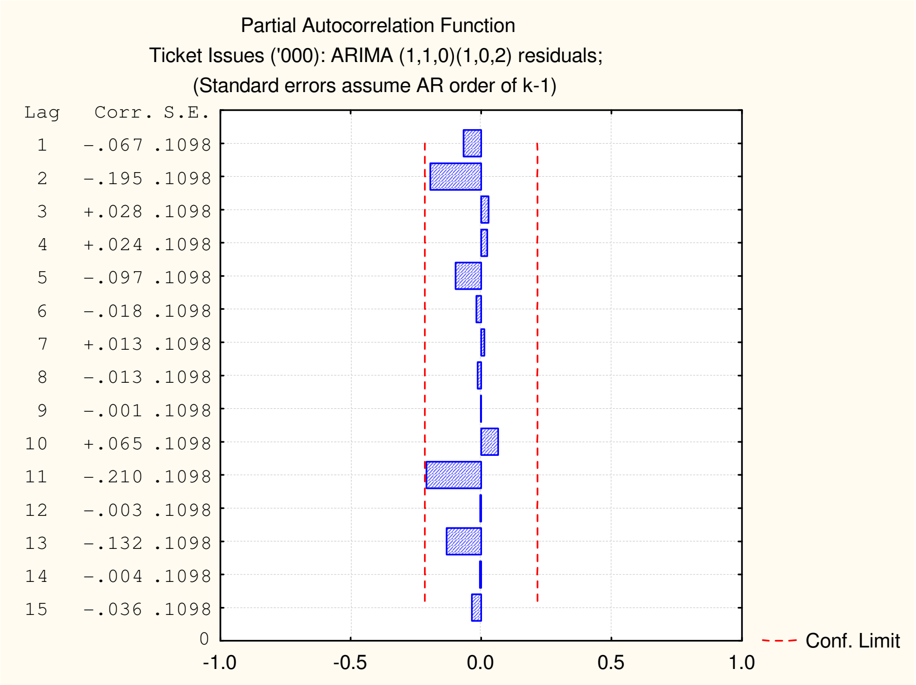
Graph 30: Autocorrelation residuals : ARIMA (1,1,0)(1,0,2)

The partial autocorrelation function with no significant partial autocorrelations is as follow:



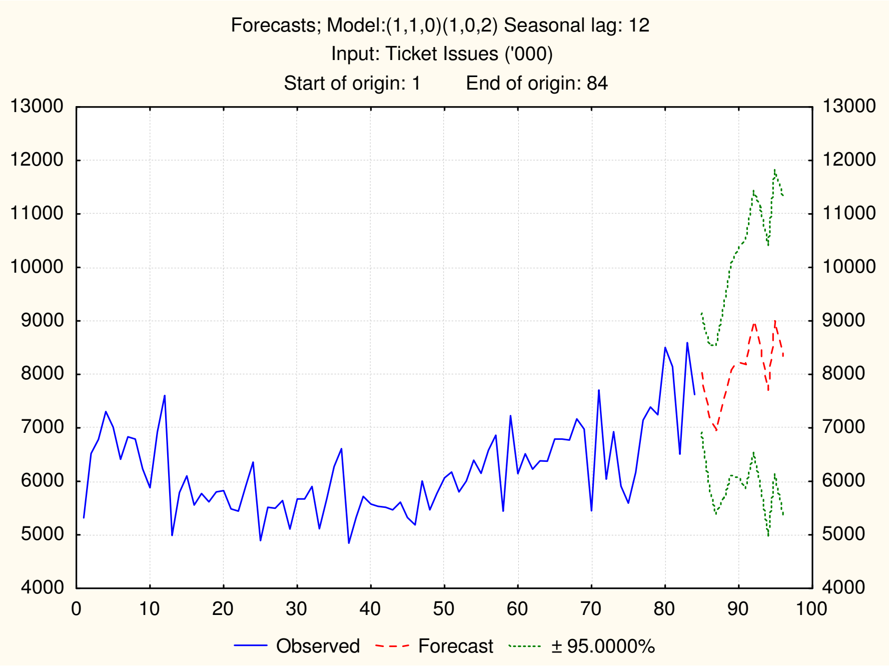
	Partial Autocorrelation Function (Ticket issues) Ticket Issues ('000): ARIMA (1,1,0)(1,0,2) residuals; (Standard errors assume AR order of k-1)	
Lag	Partial- Auto r.	Std.Err.
1	-0.0670	0.1098
2	-0.1952	0.1098
3	0.0279	0.1098
4	0.0241	0.1098
5	-0.0973	0.1098
6	-0.0185	0.1098
7	0.0127	0.1098
8	-0.0129	0.1098
9	-0.0010	0.1098
10	0.0650	0.1098
11	-0.2097	0.1098
12	-0.0030	0.1098
13	-0.1321	0.1098
14	-0.0040	0.1098
15	-0.0357	0.1098

Table 12: Partial Autocorrelation of residuals: ARIMA(1,1,0)(1,0,2)



Graph 31: Partial Autocorrelation of residuals: ARIMA(1,1,0)(1,0,2)

The forecast of this model at a 95% confidence level and a 0.05 significance level is as follow:



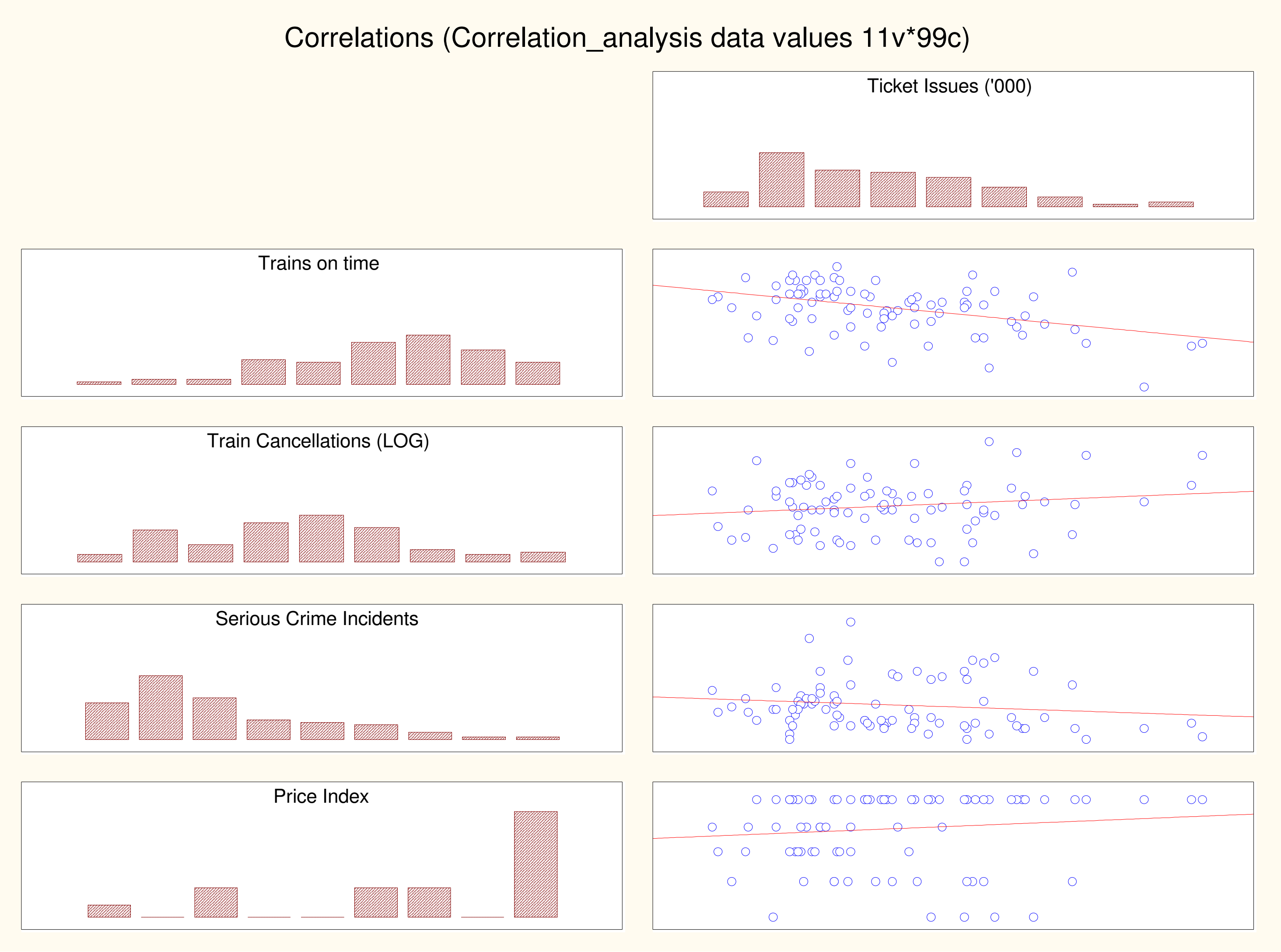
Graph 32: Forecast of ARIMA(1,1,0)(1,0,2)

The last measure of reliability of the ARIMA models applied was to compare both ARIMA models discussed with the actual ticket sales of 2007/08, to determine the best model.

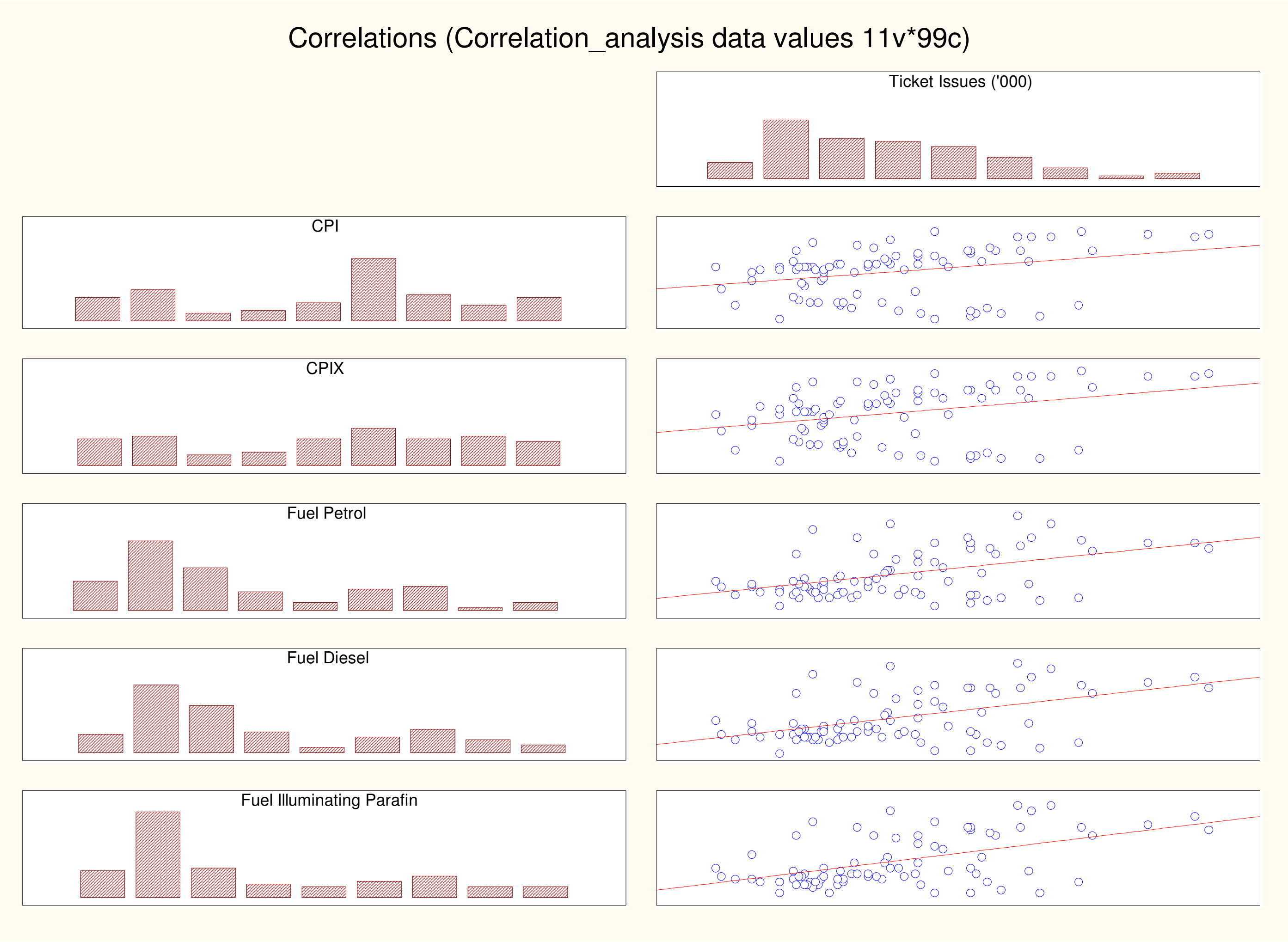
Using the data collected on independent variables, the regression model will be discussed next.

4.2.7 Regression and Multi-Regression model:

The scatter plots of the independent variables against the dependent variable ticket issues were done and are presented in Graph 33 and Graph 34.



Graph 33: Scatter plots of Ticket Issues and Internal predictors



Graph 34: Scatter plots of Ticket Issues and External predictors

The next section deals with hypothesis testing of relations between ticket issues and the independent variables selected for the study at a 90% confidence level.

4.2.7.1 Hypotheses testing:

The hypotheses tested was done at a 90% confidence level for the dependent variable ticket issues and the independent variables, trains on time, trains cancelled, serious crime incidents, price index, CPI, CPIX, fuel prices for petrol, diesel and illuminating paraffin.

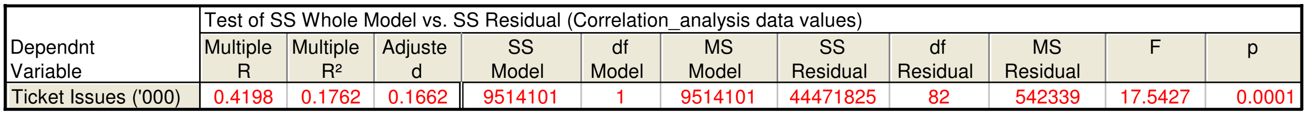
4.2.7.1.1 Hypothesis for a relation between ticket issues and trains on time

The hypothesis was as follow:

H_{01}: There is no relation between Ticket Issues and Trains-on-time

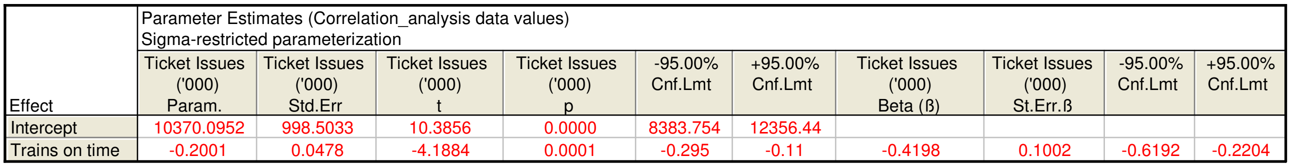
H_{11}: There is a negative relation between Ticket Issues and Trains-on-time

The one-tailed test at a 90% confidence level rejected the null hypothesis and it is concluded that there is a negative relation between trains on time and ticket issues.



Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values) Multiple R	Multiple R²	Adjusted	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.4198	0.1762	0.1662	9514101	1	9514101	44471825	82	542339	17.5427	0.0001

Table 13: Correlation results for ticket issues and trains-on-time: F-test



Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt
Intercept	10370.0952	998.5033	10.3856	0.0000	8383.754	12356.44				
Trains on time	-0.2001	0.0478	-4.1884	0.0001	-0.295	-0.11	-0.4198	0.1002	-0.6192	-0.2204

Table 14: Correlation results for ticket issues and trains on time: t-statistic

The relation is however weak as displayed by the R^2-value.

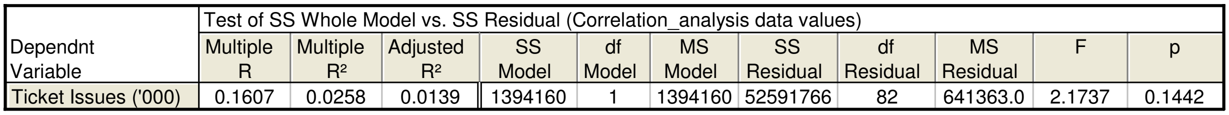
4.2.7.1.2 Hypothesis for a relation between ticket issues and trains cancelled

Due to the large variance in the trains cancelled values, the variable was transformed using a log function. The hypothesis that was tested at a 90% confidence level and 5% significance level for one-tailed test was:

H_{02}: There is no relation between Ticket Issues and Trains-cancelled

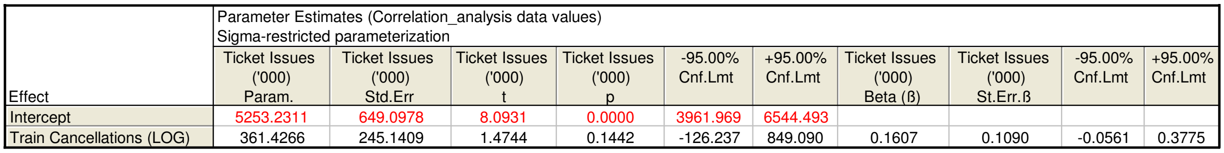
H_{12}: There is a negative relation between Ticket Issues and Trains-cancelled

The one-tailed test at a 90% confidence level indicated that there is no relation between ticket issues and trains cancelled. The null hypothesis is therefore accepted.



	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values)										
Dependnt Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.1607	0.0258	0.0139	1394160	1	1394160	52591766	82	641363.0	2.1737	0.1442

Table 15: Correlation result for ticket issues and trains cancelled: F-test



	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization									
Effect	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-95.00% Cnf.Lmt	+95.00% Cnf.Lmt
Intercept	5253.2311	649.0978	8.0931	0.0000	3961.969	6544.493				
Train Cancellations (LOG)	361.4266	245.1409	1.4744	0.1442	-126.237	849.090	0.1607	0.1090	-0.0561	0.3775

Table 16: Correlation result for ticket issues and trains cancelled: t-statistic

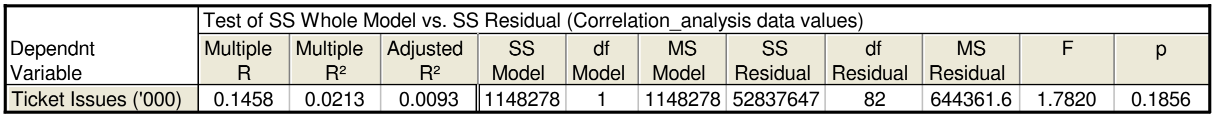
4.2.7.1.3 Hypothesis for a relation between ticket issues and serious crime incidents

The hypothesis tested was:

H_{03}: There is no relation between Ticket Issues and Serious Crime Incidents

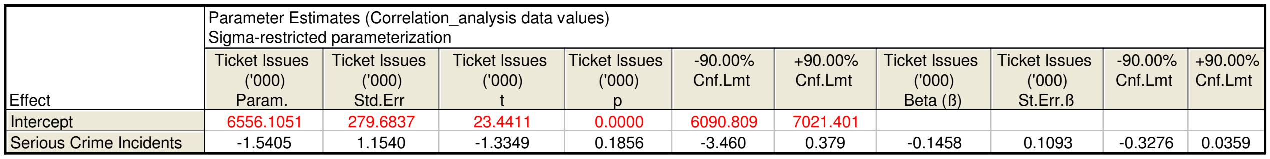
H_{13}: There is a negative relation between Ticket Issues and Serious Crime Incidents

The one-tailed test at a 90% confidence level indicated that there is no relation between ticket issues and serious crime incidents. The null hypothesis is therefore accepted.



Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values) Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.1458	0.0213	0.0093	1148278	1	1148278	52837647	82	644361.6	1.7820	0.1856

Table 17: Correlation results for ticket issues and serious crime incidents: F-test



Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	6556.1051	279.6837	23.4411	0.0000	6090.809	7021.401				
Serious Crime Incidents	-1.5405	1.1540	-1.3349	0.1856	-3.460	0.379	-0.1458	0.1093	-0.3276	0.0359

Table 18: Correlation results for ticket issues and serious crime incidents: t-statistic

4.2.7.1.4 Hypothesis for a relation between ticket issues and price index

The hypothesis tested was:

H_{04}: There is no relation between Ticket Issues and Price Index

H_{14}: There is a negative relation between Ticket Issues and Price Index

The one-tailed test at a 90% confidence level indicated that there is no relation between ticket issues and price index.

Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values)										
	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.1157	0.0134	0.0014	723217.8	1	723217.8	53262708	82	649545.2	1.1134	0.2944

Table 19: Correlation results for ticket issues and price index: F-test

Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization									
	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	5236.5397	918.7337	5.6997	0.0000	3708.089	6764.990				
Price Index	7.4751	7.0841	1.0552	0.2944	-4.310	19.261	0.1157	0.1097	-0.0667	0.2982

Table 20: Correlation results for ticket issues and price index: t-statistic

4.2.7.1.5 Hypothesis for a relation between ticket issues and inflation (CPI)

The hypothesis tested was:

H_{05}: There is no relation between Ticket Issues and CPI

H_{15}: There is a positive relation between Ticket Issues and CPI

The one-tailed test at a 90% confidence level rejected the null hypothesis and therefore there is a positive relation between ticket issues and inflation as per the CPI.

Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values)										
	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.3069	0.0942	0.0831	5084472	1	5084472	48901453	82	596359.2	8.5259	0.004520

Table 21: Correlation results for ticket issues and CPI: F-test

Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization									
	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	3583.2991	900.6316	3.9787	0.0001	2084.964	5081.634				
CPI	21.8391	7.4794	2.9199	0.0045	9.396	34.282	0.3069	0.1051	0.1320	0.4817

Table 22: Correlation results for ticket issues and CPI: t-statistic

The relationship is however weak as displayed by the R^2-value.

4.2.7.1.6 Hypothesis for a relation between ticket issues and inflation excluding interest rates (CPIX)

The hypothesis tested was:

H_{06}: There is no relation between Ticket Issues and CPIX

H_{16}: There is a positive relation between Ticket Issues and CPIX

The one-tailed test at a 90% confidence level rejected the null hypothesis and therefore there is a positive relation between ticket issues and inflation as per the CPIX.

	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values)									
Dependnt Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.3257	0.1060	0.0951	5725182	1	5725182	48260744	82	588545.7	9.7277	0.0025

Table 23: Correlation results for ticket issues and CPIX: F-test

	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization									
Effect	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	3870.4418	752.0731	5.1464	0.0000	2619.256	5121.628				
CPIX	18.8601	6.0470	3.1189	0.0025	8.800	28.920	0.3257	0.1044	0.1519	0.4994

Table 24: Correlation results for ticket issues and CPIX: t-statistic

The relationship is however weak as displayed by the R^2-value.

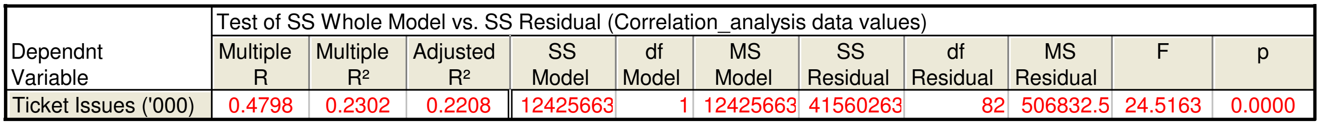
4.2.7.1.7 Hypothesis for a relation between ticket issues and fuel price for petrol

The hypothesis tested was:

H_{07}: There is no relation between Ticket Issues and Fuel price: Petrol

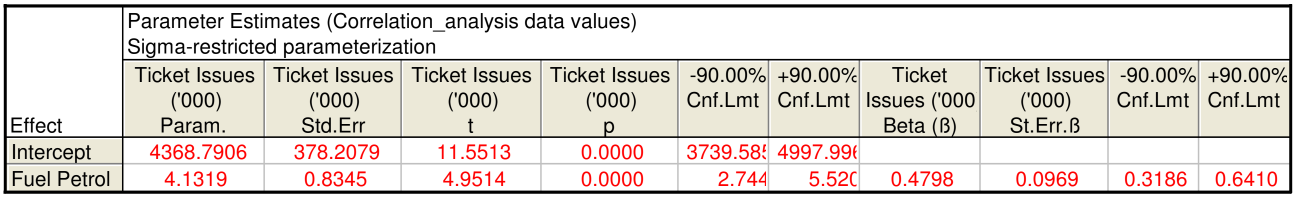
H_{17}: There is a positive relation between Ticket Issues and Fuel price: Petrol

The one-tailed test at a 90% confidence level rejected the null hypothesis and therefore there is a positive relation between ticket issues and the Fuel Price for Petrol.



Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values) Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.4798	0.2302	0.2208	12425663	1	12425663	41560263	82	506832.5	24.5163	0.0000

Table 25: Correlation results for ticket issues and price of petrol: F-test



Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000 Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	4368.7906	378.2079	11.5513	0.0000	3739.585	4997.996				
Fuel Petrol	4.1319	0.8345	4.9514	0.0000	2.744	5.520	0.4798	0.0969	0.3186	0.6410

Table 26: Correlation results for ticket issues and price of petrol: t-statistic

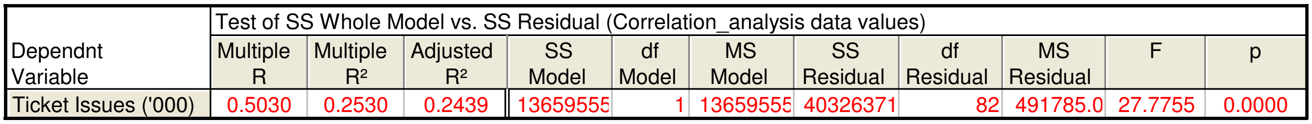
The relationship is however weak as displayed by the R^2-value.

4.2.7.1.8 Hypothesis for a relation between ticket issues and fuel price for diesel

The hypothesis tested was: H_{08}: There is no relation between Ticket Issues and Fuel price: Diesel

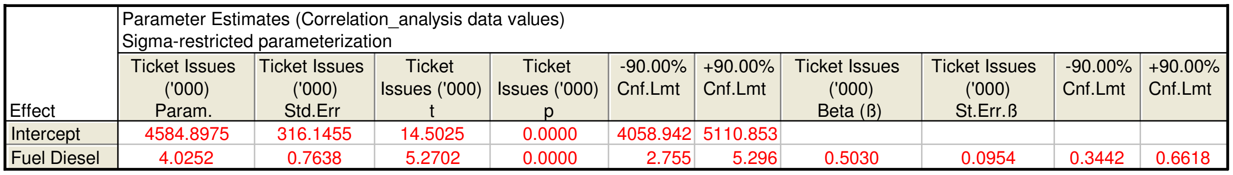
H_{18}: There is a positive relation between Ticket Issues and Fuel price: Diesel

The one-tailed test at a 90% confidence level rejected the null hypothesis and therefore there is a positive relation between ticket issues and the fuel price for diesel.



Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values) Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.5030	0.2530	0.2439	13659555	1	13659555	40326371	82	491785.0	27.7755	0.0000

Table 27: results for ticket issues and price of diesel: F-test



Parameter Estimates (Correlation_analysis data values)
Sigma-restricted parameterization

Effect	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	4584.8975	316.1455	14.5025	0.0000	4058.942	5110.853				
Fuel Diesel	4.0252	0.7638	5.2702	0.0000	2.755	5.296	0.5030	0.0954	0.3442	0.6618

Table 28: Correlation results for ticket issues and price of diesel: t-statistic

The relationship is however weak as displayed by the R^2-value.

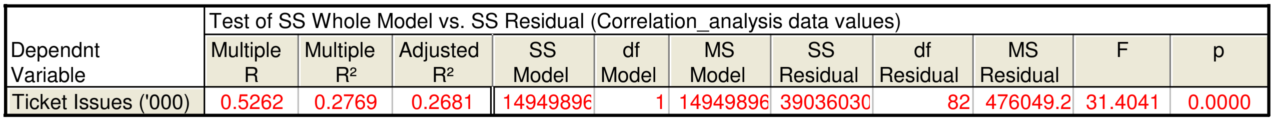
4.2.7.1.9 Hypothesis for a relation between ticket issues and fuel price for illuminating paraffin

The hypothesis tested was:

H_{09}: There is no relation between Ticket Issues and Fuel Price: Illuminating paraffin.

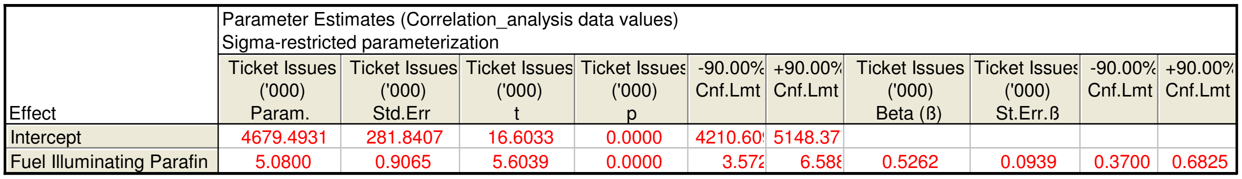
H_{19}: There is a positive relation between Ticket Issues and Fuel Price: Illuminating paraffin.

The one-tailed test at a 90% confidence level rejected the null hypothesis and therefore there is a positive relation between ticket issues and Fuel price: illuminating paraffin.



Dependnt Variable	Test of SS Whole Model vs. SS Residual (Correlation_analysis data values) Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Ticket Issues ('000)	0.5262	0.2769	0.2681	14949896	1	14949896	39036030	82	476049.2	31.4041	0.0000

Table 29: Correlation results for ticket issues and price of illuminating paraffin: F-test



Effect	Parameter Estimates (Correlation_analysis data values) Sigma-restricted parameterization									
	Ticket Issues ('000) Param.	Ticket Issues ('000) Std.Err	Ticket Issues ('000) t	Ticket Issues ('000) p	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt	Ticket Issues ('000) Beta (ß)	Ticket Issues ('000) St.Err.ß	-90.00% Cnf.Lmt	+90.00% Cnf.Lmt
Intercept	4679.4931	281.8407	16.6033	0.0000	4210.609	5148.377				
Fuel Illuminating Parafin	5.0800	0.9065	5.6039	0.0000	3.572	6.588	0.5262	0.0939	0.3700	0.6825

Table 30: Correlation results for ticket issues and price of illuminating paraffin: t-statistic

The relationship is however weak as displayed by the R^2-value.

The next section evaluates whether there are violations of assumptions for the regression model by testing multicolinearity, serial correlation and heteroscedacity.

4.2.7.2 Tests for violations of assumptions:

The tests conducted tested whether the independent variables are highly correlated (multicolinearity), whether the error terms are uncorrelated and whether the relation between the dependent variable and independent variable displays heteroscedaticity.

4.2.7.2.1 Multicollinearity:

(Hanke and Reitsch, 1998) indicates that multicollinearity occurs where more than one independent variable are too closely related to each other. The independent variables CPI, CPIX, Fuel price for petrol, Fuel price for diesel and Fuel price for illuminating paraffin were checked for multicollinearity.

The results of the tests for correlation between CPI and CPIX were as follows:

Regression Statistics	
Multiple R	0.9938
R Square	0.9877
Adjusted R Square	0.9875
Standard Error	1.2661
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10528.999	10529	6568.76534	4.73118E-80
Residual	82	131.437	1.6029		
Total	83	10660.436			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	19.9199	1.2411	16.050	5.0036E-27	17.4508	22.3889	17.8550	21.9847
CPIX	0.8088	0.0100	81.048	4.73118E-80	0.7890	0.8287	0.7922	0.8254

Table 31: Regression between CPI and CPIX

The correlation coefficient and R^2 are both high indicating high colinearity between CPI and CPIX.

The result of the test for multicollinearity between CPIX and Fuel price for petrol was as follow:

Regression Statistics	
Multiple R	0.8618
R Square	0.7427
Adjusted R Squa	0.7395
Standard Error	7.1070
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11953.6744	11953.6744	236.6644	6.85849E-26
Residual	82	4141.7355	50.5090		
Total	83	16095.4099			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	66.7540	3.7756	17.6805	1.0396E-29	59.2432	74.2648	60.4728	73.0353
Fuel Petrol	0.1282	0.0083	15.3839	6.8585E-26	0.1116	0.1447	0.1143	0.1420

Table 32: Regression between CPIX and Fuel Price for Petrol

The correlation coefficient and R^2 are both high indicating high colinearity between CPIX and Fuel prices for petrol.

The result of the test for multicollinearity between CPIX and Fuel price for diesel was as follow:

Regression Statistics	
Multiple R	0.8489
R Square	0.7207
Adjusted R Square	0.7173
Standard Error	7.4041
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11600.1209	11600.1209	211.6015	2.00044E-24
Residual	82	4495.2890	54.8206		
Total	83	16095.4099			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	76.4877	3.3379	22.9150	2.12E-37	69.8476	83.1278	70.9347	82.0408
Fuel Diesel	0.1173	0.0081	14.5465	2E-24	0.1013	0.1333	0.1039	0.1307

Table 33: Regression between CPIX and Fuel Price for Diesel

The correlation coefficient and R^2 are both high indicating high colinearity between CPIX and Fuel prices for diesel.

The result of the test for multicollinearity between CPIX and Fuel price for illuminating paraffin was as follow:

Regression Statistics	
Multiple R	0.8004
R Square	0.6407
Adjusted R Square	0.6363
Standard Error	8.3978
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10312.5389	10312.5389	146.2298	6.46582E-20
Residual	82	5782.8710	70.5228		
Total	83	16095.4099			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	83.6239	3.4304	24.3774	2.53E-39	76.7998	90.4480	77.9169	89.3308
Fuel Illuminating Paraffin	0.1334	0.0110	12.0926	6.47E-20	0.1115	0.1554	0.1151	0.1518

Table 34: Regression between CPIX and Fuel Price for Illuminating Paraffin

The test shows there is a medium relation between CPIX and Fuel prices for illuminating paraffin.

The result of the test for multicollinearity between Fuel prices for petrol and diesel was as follow:

Regression Statistics	
Multiple R	0.9882
R Square	0.9765
Adjusted R Square	0.9762
Standard Error	14.4528
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	710671.3003	710671.3003	3402.2423	1.54412E-68
Residual	82	17128.4236	208.8832		
Total	83	727799.7239			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	74.8082	6.5156	11.4815	9.541E-19	61.8467	87.7697	63.9686	85.6478
Fuel Diesel	0.9181	0.0157	58.3287	1.544E-68	0.8868	0.9494	0.8919	0.9443

Table 35: Regression between Fuel price for petrol and Fuel price for diesel

The correlation coefficient and R^2 are both high indicating high colinearity between Fuel prices for petrol and diesel.

The result of the test for multicollinearity between Fuel prices for petrol and illuminating paraffin was as follow:

Regression Statistics	
Multiple R	0.9710
R Square	0.9428
Adjusted R Square	0.9421
Standard Error	22.5356
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	686155.7243	686155.7243	1351.0895	1.03598E-52
Residual	82	41643.9995	507.8537		
Total	83	727799.7239			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	117.4793	9.2055	12.7619	3.5485E-21	99.1666	135.7920	102.1646	132.7940
Fuel Illuminating Paraffin	1.0883	0.0296	36.7572	1.036E-52	1.0294	1.1472	1.0391	1.1376

Table 36: Regression between Fuel price for petrol and Price of Illuminating Paraffin

The correlation coefficient and R^2 are both high indicating high colinearity between Fuel prices for petrol and illuminating paraffin.

Due to the high colinearity between some of the variables only CPIX and Illuminating paraffin was used in further analysis.

The next section evaluates the variables for heteroscedaticity.

4.2.7.2.2 Heteroscedaticity test

The Koenker-Basset (KB) test was used to test for heteroscedaticity between ticket issues and trains on time, ticket issues and CPIX and ticket issues and illuminating paraffin prices.

The KB test for heteroscedaticity for ticket issues and trains on time indicates that the assumption of homoscedaticity has been violated at a 90% confidence level with a t-value of 2.155.

Regression Statistics	
Multiple R	0.2315
R Square	0.0536
Adjusted R Square	0.0421
Standard Error	4189015.793
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8.15015E+13	8.15E+13	4.6445	0.0341
Residual	82	1.43892E+15	1.75E+13		
Total	83	1.52043E+15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	37888516.02	556361.2723	68.1006	6.07E-74	36781736.24	38995295.8	36962926	38814106
Residuals	1.2913	0.5992	2.1551	0.0341	0.0993	2.4833	0.2945	2.2882

Table 37: Results of KB test for trains on time

The correction applied was to use the log values of trains on time for the regression.

The KB test for heteroscedaticity between ticket issues and CPIX indicates that the assumption of homoscedaticity has not been violated at a 90% confidence level with a t-value of 0.591.

Regression Statistics	
Multiple R	0.0651
R Square	0.0042
Adjusted R Square	-0.0079
Standard Error	3254153.12
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.70128E+12	3.701E+12	0.3495	0.5560
Residual	82	8.6834E+14	1.059E+13		
Total	83	8.72041E+14			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	38372726.8	440712.4949	87.0698	1.418E-82	37496009.31	39249444.4	37639535.85	39105917.82
Residuals	0.2687	0.4544	0.5912	0.5560	-0.6353	1.1726	-0.4873	1.0246

Table 38: Results of KB test for CPIX

The KB test for heteroscedaticity for prices of ticket issues and illuminating paraffin indicates that the assumption of homoscedaticity has not been violated at a 90% confidence level with a t-value of 0.946.

Regression Statistics	
Multiple R	0.1039
R Square	0.0108
Adjusted R Square	-0.0013
Standard Error	5417102.767
Observations	84

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.62E+13	2.6242E+13	0.8943	0.3471
Residual	82	2.40629E+15	2.9345E+13		
Total	83	2.43253E+15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	38274014.45	704697.4086	54.3127	4.6319E-66	36872146.84	39675882.1	37101645.2	39446383.7
Residuals	0.7809	0.8257	0.9457	0.3471	-0.8618	2.4235	-0.5929	2.1546

Table 39: Results of KB test for illuminating paraffin

4.2.7.2.3 Serial correlation

A combination of the independent variables, trains on time, CPIX and the price of illuminating paraffin results in a Durbin-Watson statistic of 1.54. This is lower than the lower bound of 1.62 at a 0.05 significance one-sided test indicating serial correlation.

	Durbin-Watson d (Correlation_analysis data values) and serial correlation of residuals	
	Durbin-Watson d	Serial Corr.
Estimate	1.5410	0.2105

Table 40: Durbin-Watson statistic for multi-regression variables

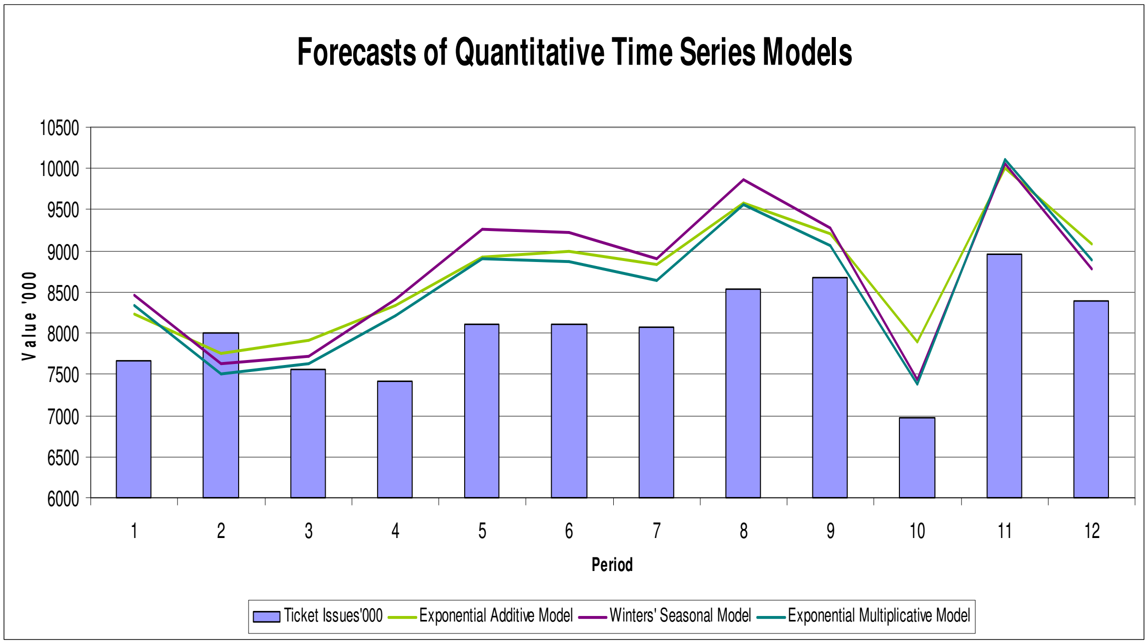
This serial correlation needs to be removed prior to evaluating the regression model for its effectiveness. Serial correlation is the result of an omitted variable or the correlation of independent error terms. The correction is thus either to find the missing variables or to create new variables through regression of percentage changes or iterative processes, such as transforming the data with lags. Transformations with lags were done for all variables and in various combinations. This has however resulted in even higher levels of serial correlation.

As the independent variables all render low relationship levels it is concluded that a multi-regression model is not suitable for forecasting of ticket issues due to the lack of sufficient explanatory variables. The conclusion is that the explanatory variables chosen only play a part in the explanation of changes in ticket issues.

The next section deals with the comparison of the various time series models found to be applicable.

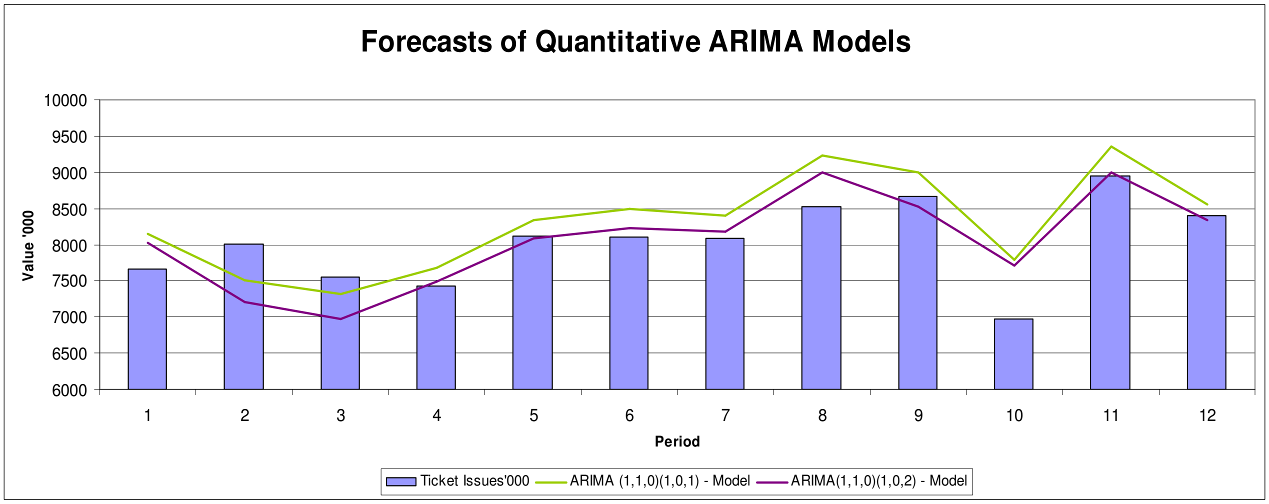
4.3 Comparison of various quantitative models applied.

The models to be compared are the Exponential Additive time series model, the Winter’s multiplicative seasonal model, the Exponential Multiplicative model and the two Box-Jenkins ARIMA models – ARIMA (1,1,0)(1,0,1) and ARIMA (1,1,0)(1,0,2) models. The forecasts of the models were graphically plotted and are depicted in Graph 35 and Graph 36.



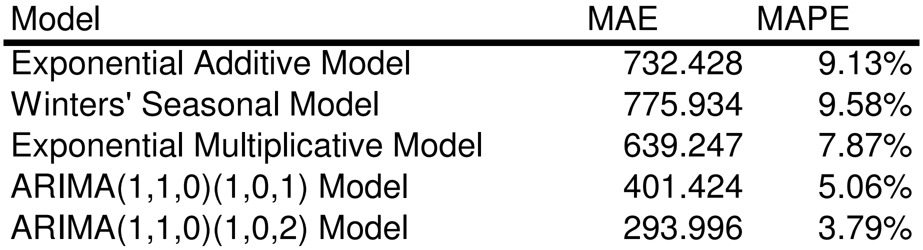
Graph 35: Comparison of forecast results for time series forecasts

The time series models all produced higher values that the actual data of 2007/08.



Graph 36: Comparison of forecast results of ARIMA forecasts

The best model would be the one that minimises the error in forecasting and to this end the residuals were compared between the models. The results of the comparisons of Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are provided in **Error! Reference source not found.**.



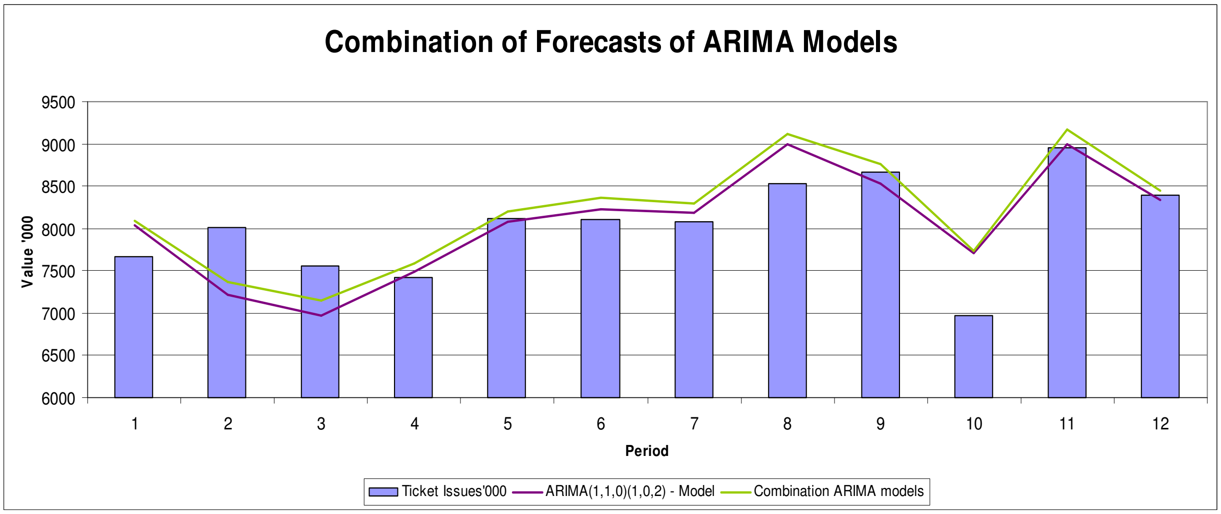
Model	MAE	MAPE
Exponential Additive Model	732.428	9.13%
Winters' Seasonal Model	775.934	9.58%
Exponential Multiplicative Model	639.247	7.87%
ARIMA(1,1,0)(1,0,1) Model	401.424	5.06%
ARIMA(1,1,0)(1,0,2) Model	293.996	3.79%

Table 41: Comparison of Error values of different quantitative forecast models

The best model with the lowest error terms are the Box-Jenkins model with a MAE of 293.996 or 3,79%.

4.4 Combination of forecasts:

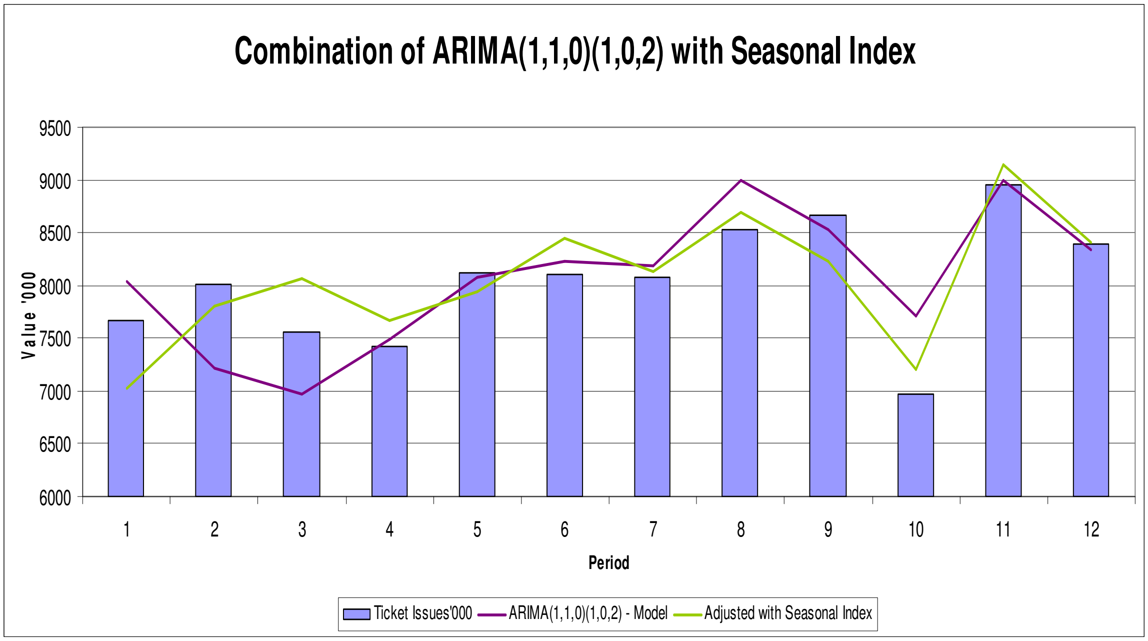
Combination of forecasts as discussed in the literature review is a way of improving the accuracy of forecasts. As the ARIMA models presented the lowest MAE and MAPE the ARIMA model forecasts were combined with equal weighting for both forecasts and compared to the actual data of 2007/08. The graphical presentation of this combination is depicted in Graph 37.



Graph 37: Comparison of combination of forecasts of ARIMA models

The MAE for this combination forecasts is 328.778 and MAPE is 4.21% indicating that the performance is not better than for the ARIMA(1,1,0)(1,0,2) model.

Another combination methodology is to overlay the quantitative forecast model with the results of the time series decomposition. The total number of ticket sales for the period 2007/08 was 96,5 million tickets and the ARIMA(1,1,0)(1,0,2) model forecasted 96,8 million tickets over the same period. This is a difference of 0.30%. As the ARIMA(1,1,0)(1,0,2) model already indicate an upward trend the seasonality index was applied. The total ticket sales forecasted by the ARIMA(1,1,0)(1,0,2) model were then adjusted with the seasonal index determined for the time series. The results for this combination provided a MAE of 24.346 and MAPE of 0.36%. This is graphically presented in Graph 38.

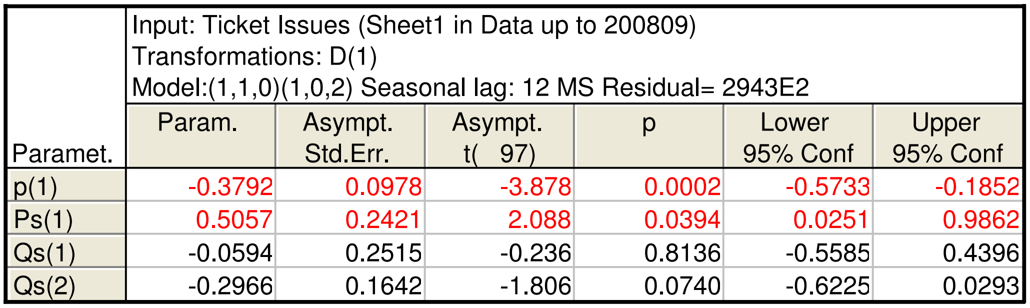


Graph 38: Combination of ARIMA(1,1,0)(1,0,2) model with seasonal index

The application of the above model to forecast the ticket sales for the period October 2008 – March 2010 is discussed in the next section.

4.5 Forecast of Ticket Sales for Commuter Rail for October 2008 to March 2010.

The comparison of the models in section 4.2 revealed that the ARIMA(1,1,0)(1,0,2) model was the best quantitative model for forecasting ticket sales. This model was then used to forecast the sales for 18 months from October 2008 – March 2010. The actual values of April 2008 – September 2008 was included in the actual data values for the ARIMA(1,1,0)(1,0,2) model. The parameters for the ARIMA(1,1,0)(1,0,2) model were as follow at a 95% confidence level:

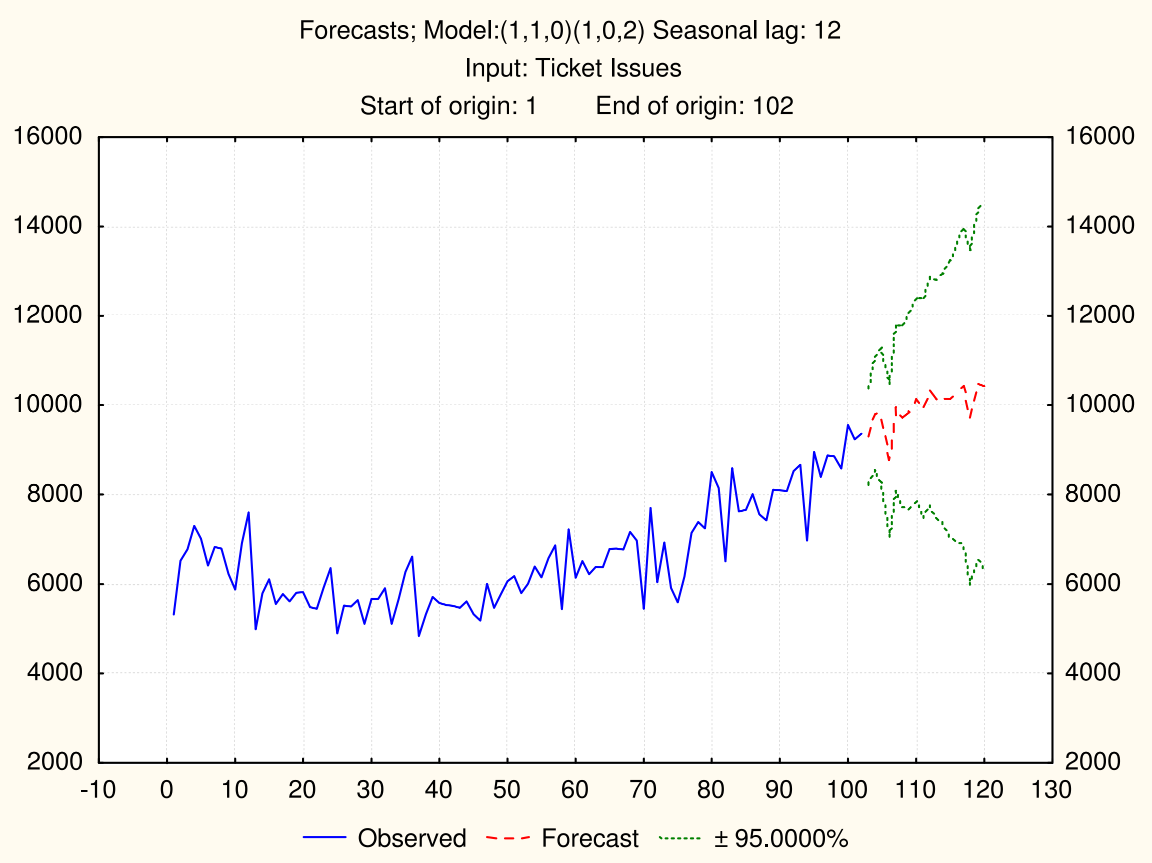


Input: Ticket Issues (Sheet1 in Data up to 200809)
Transformations: D(1)
Model:(1,1,0)(1,0,2) Seasonal lag: 12 MS Residual= 2943E2

Paramet.	Param.	Asympt. Std.Err.	Asympt. t(97)	p	Lower 95% Conf	Upper 95% Conf
p(1)	-0.3792	0.0978	-3.878	0.0002	-0.5733	-0.1852
Ps(1)	0.5057	0.2421	2.088	0.0394	0.0251	0.9862
Qs(1)	-0.0594	0.2515	-0.236	0.8136	-0.5585	0.4396
Qs(2)	-0.2966	0.1642	-1.806	0.0740	-0.6225	0.0293

Table 42: Parameters for ARIMA (1,1,0)(1,0,2) model for forecast

The forecast of this model on the data from April 2000 to September 2008 is depicted in Graph 38. The ARIMA(1,1,0)(1,0,2) model indicates a further increase in sales from October 2008 – March 2010. On a financial year basis ticket sales will increase by 16% to 111,8 million tickets for the period 2008/09 (incorporating the actual sales from April 2008 to September 2008) and by a further 9% to 122,1 million tickets for the period 2009/10. This would indicate a continued growth in fare revenue for the period as well.



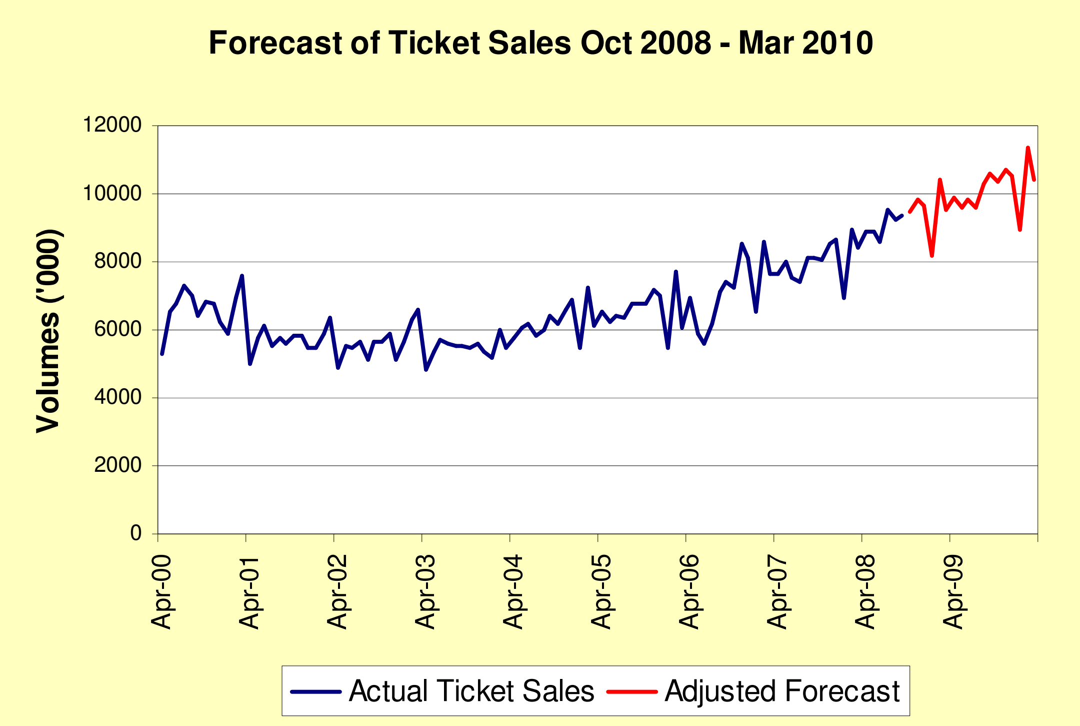
Graph 39: Forecast of ticket sales for October 2008 to March 2010

The combination of the ARIMA(1,1,0)(1,0,2) model with the seasonal index was then applied to improve the forecast. The seasonal index was adjusted with the added data from 2007/08 and 2008/09 to September 2008 to incorporate any recent changes to the seasonal pattern. The forecasted monthly data for the period October 2008 to March 2010 are as follow:

	ARIMA(1,1,0)(1,0,2) Forecast	Lower 95%	Upper 95%	Std. Error	Seasonal Index	Adjusted Forecast with Seasonal Index
Oct-08	9,298.272	8,221.573	10,374.971	542.493	1.019	9,497.525
Nov-08	9,803.404	8,536.110	11,070.698	638.524	1.055	9,829.608
Dec-08	9,781.045	8,269.838	11,292.252	761.419	1.033	9,625.516
Jan-09	8,769.560	7,075.975	10,463.145	853.310	0.877	8,169.466
Feb-09	9,950.060	8,082.640	11,817.479	940.897	1.116	10,396.838
Mar-09	9,722.250	7,699.117	11,745.383	1,019.353	1.023	9,534.892
Apr-09	9,843.534	7,674.693	12,012.375	1,092.768	0.970	9,860.674
May-09	10,142.398	7,837.451	12,447.345	1,161.344	0.945	9,607.230
Jun-09	9,934.201	7,500.601	12,367.801	1,226.166	0.969	9,850.867
Jul-09	10,319.917	7,764.185	12,875.648	1,287.702	0.944	9,604.575
Aug-09	10,134.365	7,462.058	12,806.671	1,346.438	1.010	10,267.749
Sep-09	10,166.362	7,382.365	12,950.359	1,402.713	1.040	10,579.295
Oct-09	10,136.812	7,025.501	13,248.124	1,567.630	1.019	10,367.321
Nov-09	10,305.416	6,985.464	13,625.367	1,672.753	1.055	10,729.817
Dec-09	10,434.547	6,888.525	13,980.569	1,786.658	1.033	10,507.033
Jan-10	9,720.775	5,973.125	13,468.426	1,888.249	0.877	8,917.636
Feb-10	10,486.328	6,543.464	14,429.192	1,986.607	1.116	11,348.994
Mar-10	10,424.646	6,297.202	14,552.090	2,079.607	1.023	10,408.109

Table 43: Application of ARIMA(1,1,0)(1,0,2) model with seasonal index

The forecast is graphically depicted in Graph 39.



Graph 40: Forecast of Ticket Sales using seasonal index adjustment

The conclusions and recommendations based on the research conducted and results of the study are discussed in the next Chapter.

5 Conclusions and recommendations

The objectives of the study were to:

- Determine the influence of various external factors such as economic growth or economic indicators as well as internal factors such as service delivery factors on the performance of ticket issues and
- To determine a quantitative statistical model or models that will provide the best forecasting results for ticket issues in the commuter rail environment that can be used relatively easily in the future.

The conclusions will be discussed first and then recommendations.

5.1 Conclusions from the study conducted

It was found that the factors selected for testing the influence of external and internal factors on the performance of the ticket issues, did not explain the trends in ticket issues. The following factors had relations with ticket issues at a 90% confidence level:

- Trains-on-time had a negative relation with ticket issues with a R^2 of 0.176;
- CPI had a positive relation with ticket issues with a R^2 of 0.094;
- CPIX had a positive relation with ticket issues with a R^2 of 0.095;
- Fuel prices for petrol had a positive relation with ticket issues with a R^2 of 0.230;
- Fuel prices for diesel had a positive relation with ticket issues with a R^2 of 0.253;
- Fuel prices for petrol had a positive relation with ticket issues with a R^2 of 0.277;

The relationships are all weak but fuel prices can play a role although weak in the influence of commuters to use rail.

The use of a multi-regression model could not be established due to violations of assumptions of the multi-regression model. It was found that there is high multicolinearity between CPI and CPIX and between CPIX and the fuel prices for diesel and petrol. A medium level of colinearity was found between CPIX and the price of Illuminating paraffin. The assumption of homoscedaticity was violated between ticket issues and trains on time and a correction of a log value was applied. Tests revealed that the variables selected all were serially correlated. Various transformations with lags were applied but with increased levels of serial correlation.

As the independent variables all render low relationship levels as well it is concluded that multi-regression modelling is not suitable for forecasting of ticket issues due to the lack of sufficient explanatory variables. The conclusion is also that the explanatory variables chosen only play a part in the explanation of changes in ticket issues.

Quantitative model selection started with decomposition of the ticket sales data. Ticket sales data displayed an upward trend for the period tested as well as for the period forecasted. A definite seasonal pattern was also clearly visible in the graphs. The cyclical trend could not be clearly established.

The comparison of the quantitative models rendered the following results in forecasting errors:

Model	MAE	MAPE
Exponential Additive Model	732.428	9.13%
Winters' Seasonal Model	775.934	9.58%
Exponential Multiplicative Model	639.247	7.87%
ARIMA(1,1,0)(1,0,1) Model	401.424	5.06%
ARIMA(1,1,0)(1,0,2) Model	293.996	3.79%

Table 44: Comparison of forecasting errors of quantitative models

It can be concluded that the best model for forecasting of the ticket issues is the ARIMA (1,1,0)(1,0,2) model with a Mean Absolute Percentage Error (MAPE) of 3,8%. The ARIMA models also provides the values of the upper and lower 95% confidence interval and therefore also provides an interval or cone that addresses the range of values for the forecast. The results of the forecasts of the ARIMA(1,1,0)(1,0,2) model with 95% confidence levels is presented in Appendix 8.2.

As per the literature review, combinations of forecasts as tested by (Meade, 2000) and (Landram *et al*., 2008) were conducted to establish whether the forecast accuracy could be improved. A combination of the two ARIMA models equally weighted resulted in a MAE of more than the MAE for the ARIMA(1,1,0)(1,0,2) model. A combination of the ARIMA(1,1,0)(1,0,2) model with the seasonal index applied to the forecast data improved the MAE to 24.346 or MAPE of 0.36%. It is therefore concluded that the forecast accuracy can be improved by using a combination method of the ARIMA(1,1,0)(1,0,2) model with the seasonal index from the decomposition of the time series data.

Forecasts were then compiled for the period October 2008 to March 2010. The forecasts indicate that ticket sales are expected to increase for the financial year 2008/09 by 16% whilst a further increase of 9% can be expected for 2009/10. The standard error of the ARIMA(1,1,0)(1,0,2) increases as the forecast period increases it can be concluded that the accuracy of the forecast decreases over time.

5.2 Recommendations

The main recommendation from the research study is that the organisation applies the Box Jenkins or ARIMA forecasting model in conjunction with the time series decomposition seasonal index to forecast ticket sales in the commuter rail environment to enhance business planning. The ARIMA model accommodates more recent seasonality patterns and this is further enhanced by the application of the seasonal index.

Further recommendations are:

- The forecast model needs to be evaluated for changes in accuracy and parameters need to be revisited if required, to ensure that the model remains appropriate;
- Re-forecasting should be conducted on a regularly basis at least bi-annually as the standard error increases over time therefore indicating less accuracy the longer the time period;
- Re-forecasting also need to be event driven that is in cases where specific environmental changes or internal delivery issues have changed significantly it is advisable to re-do the forecasts and
- This quantitative forecast is a baseline forecast and can be improved with judgment overlay by Senior Management.

5.2.1 Recommendations for further study:

The study covered the overall results for the business. The following enhancements to the study are recommended:

- Developing an ARIMA model for the ticket sales of each region to ensure that regional differences are taken into account;

- Explore the differences in the various ticket products namely single, weekly and monthly tickets and forecasting these with the appropriate time series or ARIMA model should be investigated; and
- Development of other explanatory variables to explore the influences of internal and external factors on ticket sales. Such variables could be of a more qualitative nature or variables such as unemployment and labour force changes or the development of a price differential indicator of the price of commuter rail with other modes of public transport. These factors should be explored as it could assist in predicting future inflection points that impact the choice of rail as transport mode.

6 Acknowledgements:

I would like to acknowledge the following people and institutions for their support of this study:

- My husband, Riaan and sons, Mornè and Etienne for their continued support, motivation and patience over the last three years;
- My mother for her unfailing support and editing of the report;
- My group members for their support and friendship;
- The South African Rail Commuter Corporation for the use of company information in the conduct of this study as well as for the bursary that enabled me to do this course;
- My study leader, Dr Jan Kruger for his comments and guidance and
- Above all my God for His wonderful grace over this period.

7 References

Bickel R, 2007, *Multilevel Analysis for Applied Research: It's Just Regression*, New York: The Guilford Press

Diamantopolis A and Sclegelmilch BB, 2000, *Taking the Fear out of Data Analysis*, London: Thomson Learning

Department of Mineral and Energy, 2008, "Fuel prices" available from [www.dme.gov.za][accessed 17 August 2008]

Fader PS and Hardie BGS, Jul.-Oct. 2005, "The value of simple models in new product forecasting and customer-base analysis", *Applied Stochastic Models in Business and Industry*, 21(4/5):461-473.

Ferrara L and Guègan D, Dec. 2001, "Forecasting with *k*-factor Gegenbauer Processes: Theory and Applications"; *Journal of Forecasting*, 20(8):581-601.

Flores BE and Wichern DW, Sep. 2005, "Evaluating forecasts: A look at aggregate bias and accuracy measures", *Journal of Forecasting,* 24(6):433-451.

Foster J, Barkus E and Yavorsky C, 2006, *Understanding and Using Advanced Statistics,* London: Sage Publications Ltd.

Georgoff DM and Murdick RG, Jan.–Feb. 1986, "Manager's guide to forecasting", *Harvard Business Review*, 64(1):110-120.

Gaither N and Frazier G, 2002, *Operations Management,* 9th Edition, Ohio: South-Western a division of Thomson Learning.

Hanke JE and Reitsch AG, 1998, *Business Forecasting,*6th Edition, New Jersey: Prentice-Hall

Hill T and Lewicki P, 2006, *Statistics: Methods and applications, A comprehensive reference for science, Industry and Data Mining,* Tulsa: Statsoft, Inc.

Jain, CL, Winter 2007/2008, "Benchmarking Forecasting Process", *Journal of Business Forecasting*, 26(4): 9-23

Jain, CL, Winter 2007/2008, "Benchmarking Forecasting Models", *The Journal of Business Forecasting*, 26(4): 15-35

Jiang W, Au T, Kwok-Leaung Tsui, March 2007, "A statistical process control approach to business activity monitoring", *IIE Transactions,* 39(3): 235-249

Kedem B and Fokianos K, 2002, *Regression Models for Time series Analysis,* New Jersey: John Wiley & Sons, Inc.

Landram FG, Pavur RJ and Alidaee B, January 2008. "Combining Time-Series Components for Improved Forecasts", *Decision Sciences Journal of Innovative Education,* 6(1): 197-204.

Lapide L, Fall 2007, "How often to forecast", Journal of Business Forecasting, 26(3): 18-20.

Makridakis S and Hibon M, May 1997, "ARMA Models and the Box-Jenkins Methodology", *Journal of Forecasting,* 16(3):147-163.

Makridakis S, Wheelright SC and Hyndman RJ, 1998, *Forecasting Methods and Applications*, 3rd Edition, New York: John Wiley & Sons Inc.

Meade N, 2000, "Evidence for the Selection of Forecasting Methods", *Journal of Forecasting*, 19: 515-535

Price DHR and Sharp JA, Oct.-Dec.1988, "The impact of the performance of growth curve models of changes in parameter re-estimation period and other factors", *Journal of Forecasting,* 7(4):245-258.

Proietti T, Jan. 1998, "Seasonal heteroscedasticity and trends", *Journal of Forecasting,* 17(1):1-17.

Pycraft M, Singh H, Phihlela K, Slack N, Chambers S, Harland C, Harrison A and Johnston R, 2003, *Operations Management*, South African Edition, South Africa: Pearson Education.

Ramanujan S and Fisher A, Fall 2006, "Forecasting and planning in an extremely seasonal business – Intuit's Experience", *Journal of Business Forecasting,* 25(3):11-16;

Sadownik R and Barbosa EP, May 1999, "Short-term forecasting of industrial electricity consumption in Brazil", *Journal of Forecasting,* 18(3):215-224;

Saffo P, Nov. 2007, "Six rules for effective forecasting", *Harvard Business Review*, 85(11): 122-131.

SARCC, August 2007, "Customer Satisfaction Index – Survey 1/2007", Johannesburg

SARCC, 2006, "Commuter Profiles - 2006", Johannesburg

Statistics South Africa, 2008, "Key Indicators: CPI and CPIX" available from [www.statssa.gov.za/keyindicators/keyindicators.asp] [accessed 16 August 2008]

Steel SJ and Uys DW, Jul/Aug 2007, "Variable selection in multiple linear regression: The influence of individual cases", *ORiON*, 23(2):123-136

Tay AS and Wallis KF, Jul. 2000, "Density Forecasting: A survey", *Journal of Forecasting,* 19(4): 235-254.

Timm NH, 2002, *Applied Multivariate Analysis,* New York: Springer-Verlag Inc.

Tripodis Y and Penzer J, April 2007, "Single-season Heteroscedaticity in Time Series", *Journal of Forecasting,* 26(3):(189-202)

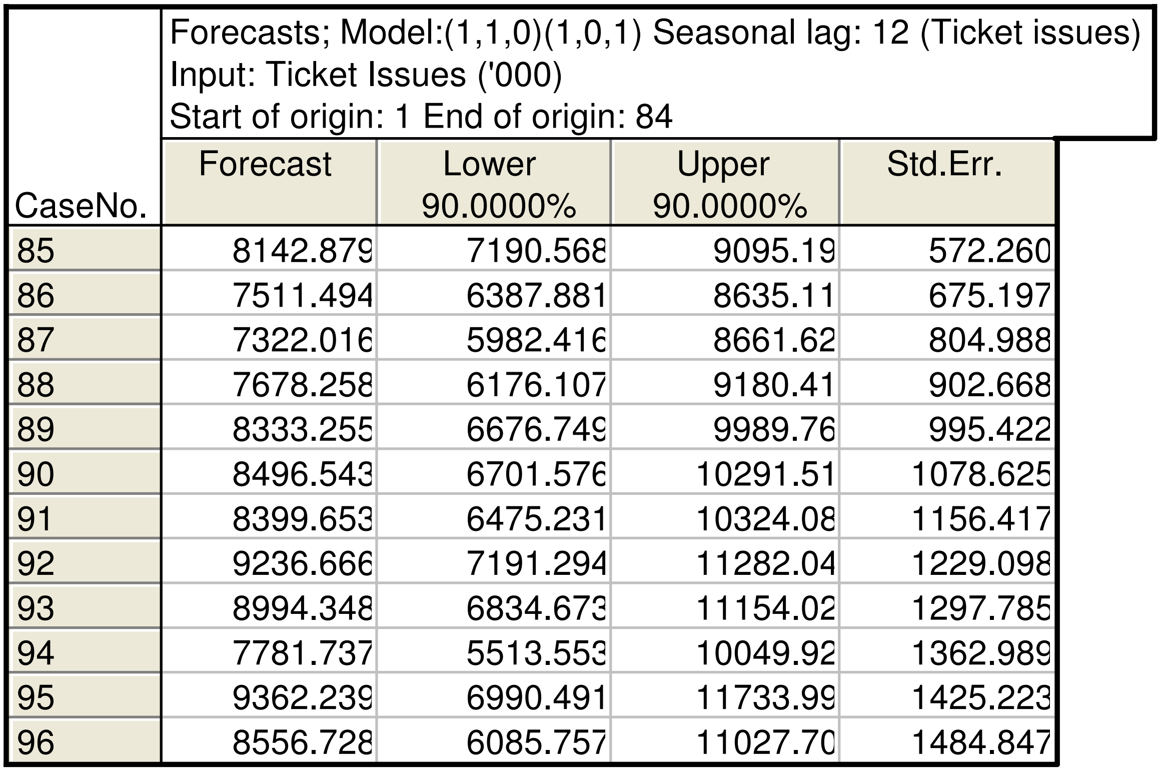
Van den Bergh F, Holloway J, Pienaar M, Koen R, Elphinstone CD, Woodborne S, Jun/July 2008, "A comparison of various modelling approaches applied to Cholera case data", *ORiON*, 24(1): 17-36.

Wang GCS, Spring 2008, "A guide to Box-Jenkins Modeling", *The Journal of Business Forecasting,* 27(1): 19-28

York D, Summer 2005, "Re-forecasting: How often for best decisions and efficiency", *Journal of Business Forecasting,* 24(2):20-23

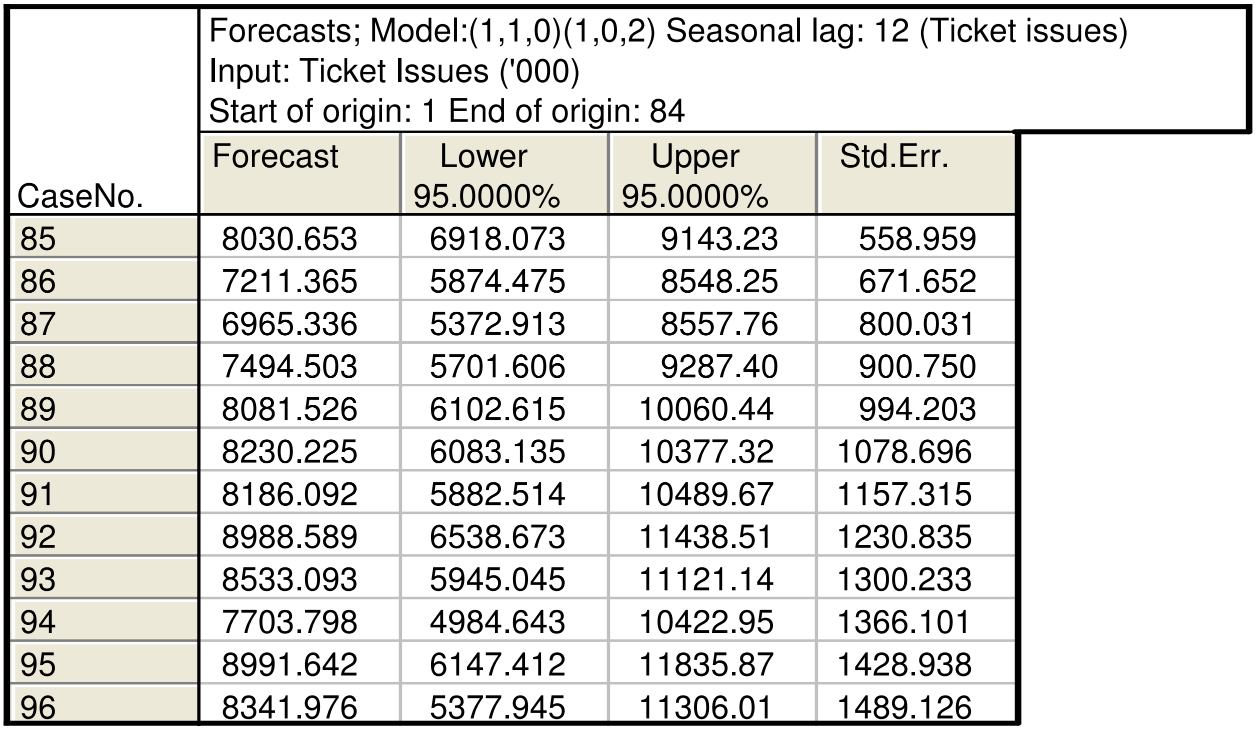
8 Appendices:

8.1 ARIMA(1,1,0)(1,0,1) model forecasts with confidence intervals



	Forecasts; Model:(1,1,0)(1,0,1) Seasonal lag: 12 (Ticket issues) Input: Ticket Issues ('000) Start of origin: 1 End of origin: 84			
CaseNo.	Forecast	Lower 90.0000%	Upper 90.0000%	Std.Err.
85	8142.879	7190.568	9095.19	572.260
86	7511.494	6387.881	8635.11	675.197
87	7322.016	5982.416	8661.62	804.988
88	7678.258	6176.107	9180.41	902.668
89	8333.255	6676.749	9989.76	995.422
90	8496.543	6701.576	10291.51	1078.625
91	8399.653	6475.231	10324.08	1156.417
92	9236.666	7191.294	11282.04	1229.098
93	8994.348	6834.673	11154.02	1297.785
94	7781.737	5513.553	10049.92	1362.989
95	9362.239	6990.491	11733.99	1425.223
96	8556.728	6085.757	11027.70	1484.847

8.2 ARIMA (1,1,0)(1,0,2) model forecasts with confidence intervals



Forecasts; Model:(1,1,0)(1,0,2) Seasonal lag: 12 (Ticket issues)
Input: Ticket Issues ('000)
Start of origin: 1 End of origin: 84

CaseNo.	Forecast	Lower 95.0000%	Upper 95.0000%	Std.Err.
85	8030.653	6918.073	9143.23	558.959
86	7211.365	5874.475	8548.25	671.652
87	6965.336	5372.913	8557.76	800.031
88	7494.503	5701.606	9287.40	900.750
89	8081.526	6102.615	10060.44	994.203
90	8230.225	6083.135	10377.32	1078.696
91	8186.092	5882.514	10489.67	1157.315
92	8988.589	6538.673	11438.51	1230.835
93	8533.093	5945.045	11121.14	1300.233
94	7703.798	4984.643	10422.95	1366.101
95	8991.642	6147.412	11835.87	1428.938
96	8341.976	5377.945	11306.01	1489.126