2022

A Comparative study of Hybrid Neural Network and ARIMA Models with application to forecasting intra-day child-line calls in Kenya.

Wang'ombe, Grace Wairimu
Strathmore Institute of Mathematical Sciences
Strathmore University

A Comparative Study of Hybrid Neural Network and ARIMA Models with Application to Forecasting Intra-day Child-line Calls in Kenya.

Grace Wairimu Wang'ombe

Submitted in total fulfilment of the requirements for the degree of Masters of Science in Statistical Science of Strathmore University

Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

October 2022

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: .... **Grace Wairimu Wang'ombe** ....

Signature: ....

Date: .... August 25, 2022 ....

Approval

The thesis of Grace Wairimu Wang'ombe was reviewed and approved by the following:

Dr. Collins Ojwang' Odhiambo
Supervisor,
Institute of Mathematical Sciences, Strathmore University.

Dr. Thaddeus Wandera Egondi
Supervisor,
Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu
Dean,
Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo
Director,
Office of Graduate Studies, Strathmore University.

Abstract

Background For successful staffing and recruiting of call centre professionals, precise forecasting of the number of calls arriving at the centre is crucial. These projections are needed for various periods, both short and long-term. Benchmark time series models such as ARIMA and Holt-Winters used in forecasting call centre data are outperformed in long term forecasts, especially when the data is not stationary. Advanced models such as the ANNs can pick up on the random peaks or outlying periods better than the benchmark time–series models. The hybrid methodology combines the strengths of the benchmark time–series and advanced models, thus improving overall forecasts.

Objective: The study's primary goal was to assess the superiority of a Hybrid ARIMA-ANN model over its constituent models in forecasting Childline call centre data in Kenya.

Methods: The ARIMA, ANN and hybrid ARIMA-ANN models were used in the call centre data forecasting. The cross-validation technique was used to create forecasting accuracy metrics which are then compared. In ARIMA, the Box-Jenkins methodology is used to fit the model whereas the neural network element of the hybrid model and the ANN were modelled using the feed-forward Neural Network Autoregressive(NNAR) structure.

Results: The Seasonal ARIMA - ANN model outperformed the ARIMA model in short term forecasts and the ANN model in long term forecasts. The Diebold-Mariano test indicated a significant difference between the hybrid and ANN forecasts, whereas the difference between the hybrid and ARIMA forecasts was not significant.

Conclusion: The Hybrid model was able to adapt both of its constituent models' advantages to better its performance. These results are helpful as call centres can be able to use one model which is robust enough to create accurate forecasts rather than the benchmark models.

Key words: call centre forecasting; benchmark methods; advanced methods; hybrid methods; forecasting horizons;

Table of contents

List of figures

List of tables

List of abbreviations

ARIMA	Auto Regressive Integrated Moving Average	ANN	Artificial Neural Networks
SSA	Singular Spectral Analysis	AIC	Akaike Information Criterion
NN	Neural Networks	RMSE	Root Mean Squared Error
ES	Exponential Smoothing	NNAR	Neural Network Auto Regression
KPSS	Kwiatkowski Phillips Schmidt Shin	ADF	Augmented Dickey Fuller
HW	Holt Winters	ACF	Auto-correlation Coefficient
PACF	Partial Auto-correlation Coefficient	SARIMA	Seasonal Auto Regressive Integrated Moving Average
TBATS	Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components	EDA	Exploratory Data Analysis
MAPE	Mean Absolute Percentage Error	BIC	Bayesian information criterion
FGM	Female Genital Mutilation	BIC	Bayesian information criterion

Acknowledgement

My sincere thanks go to Dr Collins Odhiambo and Dr Thaddeus Egondi for their noble guidance and valuable suggestions. I would also like to thank my friends and family for encouraging, proof reading and supporting me whenever I needed them.

Chapter 1

Introduction

1.1 Background of the study

Call centres are offices that handle many phone calls to respond to customer's needs. Call centre data contains inbound calls received over time. Models such as Exponential Smoothing models, ARIMA, Neural Networks, Spectrum Analysis, dynamic harmonic regression have been used in forecasting call centre data. Different studies on call centres have used parametric models such as; ARIMA and Exponential smoothing models in Taylor (2008) and non-parametric models such as artificial neural networks in Huang et al. (2019) and singular spectrum models in (Al-Azzani et al., 2020). ARIMA and Exponential Smoothing models are parametric models which assume linearity, normality or stationarity of the data. In contrast, non-parametric models such as the frequency domain models and Neural networks have no such assumptions (Al-Azzani et al., 2020), (Zhang, 2003).

Call centres are set up to take calls, some offering assistance such as emergency hotlines, ambulance services, financial centres. A call centre's capability is mainly controlled by the human resources used. Because these are costly, call centre quality is frequently matched with capacity such as staffing so that the call centre can deliver the best service at the lowest cost. Incorporating an ideal staffing level is critical; overstaffing incurs unnecessary operating expenses, whereas understaffing causes consumers to perceive queuing times as unsatisfactory, culminating in call abandonment. Staffing is, in fact, a balancing act between service quality and operational costs.

The accurate forecasting of the volume of calls arriving at the centre is critical for efficient staffing and recruitment of call centre personnel. These forecasts are required for various periods, either short-term or long-term horizons. According to Gans et al. (2003), a good strategy is to prepare an initial timetable several weeks ahead of time, then have it periodically changed until the designated day arrives. Call centres frequently track call volume throughout the day to enable dynamic real-time updating of agent deployment. Agents who become available at a moment's notice are allocated to meetings and training.

Gans et al. (2003) emphasise forecasting of arrival rates as one of the most critical areas in their projections of future directions for call centre research. Fildes and Kumar (2002) also emphasises the need for more study into time series forecasting methods in the broader telecommunications field.

In empirical investigations with intraday data, exponential smoothing has performed well in short term horizon forecasting, such as in (Taylor, 2008). The model is particularly advantageous as the researcher adds weights to the model, which determines the impact of past observations on the forecasting. The autoregressive (AR), moving average (MA), and ARIMA models are well-known statistical predictive models that forecast time-series results expressed as a mathematical function of previous numbers and white noise elements. These models perform very well when the data has a discernable pattern, i.e. trend and has very few outliers, which is not the case for most call centre data.

Data from call centres are unique and dynamic as it varies by circumstance, meaning that it can include significant effects from holidays, special events, advertisements, and unexplained fluctuations. Model formulation and parameter estimation are affected, and modelling call centre data may necessitate data purification. When outliers are caused by seasonal or advertisements, they can frequently be identified through experience and market data. Outliers that are caused by system failures or other inexplicable events, on the other hand, are more challenging to recognise and model (Barrow and Kourentzes, 2018). For this type of data, non-parametric models like Artificial Neural Network Models (ANNs) are better. Neural Networks have a great learning potential, which makes them powerful predictors. However, despite evidence that ANNs can handle the complicated structure of call centre data, there is limited research in this area.

In addition, integrated and hybrid models have been utilised in call centre data. The hybrid model combines the benefits of its constituent models. An example by Barrow (2016) assesses the utilisation of several univariate time series models in forecasting intraday call data with the data having outliers. The authors combine the Seasonal Moving Average model with the ANN model to enhance short-term accuracy without compromising long-term performance.

This project will be an empirical investigation of univariate time series methods in forecasting intraday call centre data over different horizons to show the predictive power of each model for short term and long term horizons.

1.2 Child Abuse Helpline in Kenya

Child abuse has long been documented in several studies, art and research in many regions of the world, and it continues to be a complex social problem everywhere. Child abuse can be broadly classified into physical abuse, sexual abuse, and verbal/ psychological/ emotional abuse. Under the Children Act of 2002, the government pledges to gradually use all available resources to fully realize a child's rights. The right to life, parental care, education, and health care are among these rights.

A telephone survey was conducted on 625 callers in a Childline Kenya study. Before the age of 18, 35% of the respondents had been sexually abused. The most vulnerable age for the onset of child sexual abuse was between 11 and 15 years, with an average of 13.3 years (ChildlineKenya, 2012). According to the Ministry of Labour and Social Protection's 2019 Violence Against Children Survey, one in every two young adults in Kenya had experienced violence as a kid. According to the study, 46% of 18 to 24-year-old women and 52% of young men in the same age range experienced at least one sort of violence: physical, emotional, or sexual during their youth. Child marriage is common among girls, accounting for 23% of marriages.

Since 2007, the government of Kenya has entrusted Childline Kenya, an organisation protecting children from abuse and neglect, to run the only 24-hour national child helpline in Kenya through the Department of Children Services (DCS) in the Ministry of Gender, Children and Social Development. It would create a plethora of data on child abuse, violence, and counselling needs, with far-reaching implications for child safety and national policymaking.

Childline Kenya collaborates with the government to prevent child abuse and ensure that all children live in a safe environment. They provide a 24-hour toll-free helpline dedicated to children, which may be reached by calling 116 (ChildlineKenya, 2022). Mtoto News is a digital enterprise that uses technology to improve the well-being of children. Mtoto News was founded in February 2017 to fill a gap in the online conversation about children that could lead to change. Since then, it has developed tools that help children and adults better respond to children's needs and put children at the centre of solutions through child participation (MtotoNews, 2022).

A child helpline is a phone service that links children in need of care and protection to essential services and resources. A crisis desk was set up where public members reported child abuse—members of the public call the helpline with different issues requiring different interventions. The tried cases include child neglect, physical abuse, sexual abuse, school-related, custody and maintenance, FGM, early/forced marriages, child labour, child prostitution and child trafficking. The calls are classified into the following broad categories: Abuse, education & career, Family and community, General information, Health/basic needs and Non-intervention.

1.3 Statement of the problem

Call arrival projections must be accurate to ensure optimum operating effectiveness. Since call centres data is affected by external factors such as bank holidays and school holidays, it is rarely linear. Traditional time series models require linearity, stationarity and normality assumptions to be met; thus, data transformation would be necessary before modelling. In transforming the data to meet the premises, crucial information from the data may be lost. Non-parametric forecasting techniques avoid this since they do not require these assumptions to be completed. However, the non-parametric models have previously fallen short when forecasting short term, whereas parametric models do not perform well long-term. A hybrid of the parametric and non-parametric model can improve accuracy by taking advantage of each model's strengths. This research will test whether using a hybrid forecasting technique improves call centre data prediction accuracy over constituent models.

1.4 Objective of the study

1.4.1 General Objective

This project aims to assess the performance of the hybrid forecasting model in improving the accuracy of forecasts in childline data for proper management and response of the childline centre.

1.4.2 Specific Objectives

1. To assess the performance of the hybrid forecasting model over its constituent models in forecasting the call centre data.

2. To evaluate the impact of the hybrid model over the ANN model on short term forecasts and over ARIMA model on long term forecasts.

1.4.3 Research Questions

1. Does the hybrid of the ARIMA and ANN model perform better than its constituent models?

2. How much does the hybrid model outperform the ARIMA model on short term forecasts?

3. How much does the hybrid model outperform the ANN model on long term forecasts?

1.5 Justification of the study

The aim of any statistical research is typically to estimate or infer an outcome of interest. Even for a single product or service, such as a medical emergency hotline, precise forecasting of inbound calls can have significant socio-economic ramifications. As a result, models that are best matched to the data are required. Developing relatively advanced models for estimating inbound call arrivals has recently focused on time series forecasting research.

Traditional ARIMA models have been hard to beat when it comes to short term accuracy in studies such as Barrow and Kourentzes (2018), ANN model had the best accuracy in outlying periods. However, the Traditional Model - ARIMA outperformed it in other periods. Similarly, in the study Taylor (2008), traditional models outperformed the parametric models short term, with the advanced models being competitive in the later periods of forecasts. The author then suggests that it would be prudent to combine competitive models to improve results.

Due to each model's different capabilities, there have been different studies coming up with hybrid models to combine the strength of each method (Barrow, 2016). These models have performed better than the basic ARIMA model. Call centre data has been modelled with these hybrid models in developed countries but not developing countries. As a result, this study is the first to improve the forecast of call centre data using a hybrid ARIMA- ANN model in Kenya.

1.6 Significance of the study

This study aims to identify an alternative to traditional time series models to improve call centre data prediction accuracy. The model shall then be used for accurate call centre forecasts, enabling efficiency in staffing and resourcing.

Chapter 2

Literature Review

2.1 Introduction

Different models have been used in forecasting call centre data. They include Time series models, Neural Networks and hybrid models. This section will review the Parametric time series models: Exponential smoothing and ARIMA model, Non-Parametric Models: Neural networks, Singular Spectrum Analysis and the hybrid of neural networks and ARIMA models.

2.2 Parametric Models

2.2.1 Exponential Smoothing Model

Exponential smoothing is a technique for forecasting univariate time series data. This method generates forecasts that are weighted average values of observed data, with the weights of older observational data decreasing exponentially. Exponential smoothing is a convenient procedure with no structured framework, an extension that manages developments and the most robust technique that includes seasonality support. Statisticians can alter how rapidly older findings lose their significance in the calculations by adjusting model parameters (Taylor, 2010). As a result, analysts can adjust the comparative importance of new observations versus older observational data to meet the needs of their subject area. The model is limited with the number of seasons and cycles it can accommodate; since call centre data has multiple peaks, extending the model to fit its unique characteristics is needed. Taylor (2008) extends the model so that multiple peaks can be modelled.

2.2.2 Autoregressive Moving Average (ARIMA)

ARIMA is a statistical analysis model that employs time-series data to understand the dataset and forecast future trends better. Autoregressive statistical models help predict previous values (Lin et al., 2017). This technique is used in forecasting to account for a sequence of development in the data; thus, the "auto-regressive" part, the rate of change of the growth/decline in the data, hence the "integrated part" and noise among consecutive points of time, thus the "moving average part". The autoregressive (AR), moving average (MA) and ARIMA mathematical predictive models are well-known statistical predictive models that forecast time-series findings defined as a mathematical function of previous numbers and white noise elements (Lin et al., 2017). As a result, these models impose an inherent linearity constraint on the information generation process.

Taylor (2008) compares ARIMA, Holt-Winters Exponential, Harmonic regression and periodic AR models for intraday calls for a bank in the UK. Using their prior study, the authors extend the Holt-Winters model to two seasonal cycles. Periodic AR was used as it allowed for variations across different cycles. Holt-Winters extension indicated great potential for forecasting two to three days in advance. ARIMA three to five days in advance and the basic historical average for longer durations. They compare the techniques' short-term prediction results to alternative ARIMA models that incorporate additional seasonal variables to account for the weekly cycle and conclude that the double seasonal exponential smoothing strategy is superior. The authors indicate the simplicity of exponential models over ARIMA modelling but note that robust models would be helpful to non-normal distribution data and outlying observations. The simplicity of the exponential models allows for the process's automation as the outliers were smoothed out using the historical average. In the study, exponential smoothing and ARIMA models outperformed the harmonic regression technique for short term forecasts. The authors propose that other researchers combine forecasting methods using weighted averages in situations where there are two methods with similar results as the case of the ARIMA and ES model for the case study.

2.3 Non Parametric Models

2.3.1 Frequency domain models.

The frequency-domain approach relates periodic variations caused by some phenomenon of interest, i.e. regression of sinusoids. Most Time series data have periodic aspects that cause the forecasting techniques to be inaccurate when only trend and seasonality are accounted for in the model. Frequency domain models consider cyclical time-series change. They are useful in predicting the weather and traffic; however, they are not yet widely used in many areas within phenomena that show cyclic tendencies. Only a few publications have used frequencies to forecast the call centre data.

Al-Azzani et al. (2020) did an empirical study and compared ARIMA, Holt-Winters, Multiple Linear Regression, and Single Spectrum Analysis to research which model would have superior accuracy in modelling short term and long term periods. SSA is considered flexible to model real-world scenarios as it does not require any transformations to be done, thus no loss of data. The results from the empirical study showed that ARIMA ignores the random peaks and thus ends up having less accurate results in the long term demand but better results in the short term demand.

Taylor (2008) added dynamic harmonic regression in the comparison of the univariate models. The models use trigonometric regressors with time-varying parameters. The dynamic harmonic regression model was designed for use with intraday contact centre time series. The basic dynamic harmonic regression model presented thus far was incapable of accurately modelling the auto-correlation in their series. This was also true for the original study series. To address this issue, they equipped a separate dynamic harmonic linear regression for every day of the week to the residuals of the basic model. They treated all residuals for just any given weekday as a single series and then assessed a dynamic harmonic regression model for that series. The model performed poorly short term but would compete with the other models long term. However, the data was already stationary; thus, the parametric assumptions were met.

2.3.2 Neural networks

A neural network is a network that attempts to simulate the neurons or brain tissue found in the human brain. It comprises several "nodes" that attempt to mimic the human brain's capacity. Forecasting high-frequency call arrivals obtained in hourly or smaller time buckets remains a crucial problem for call centres (Hill et al., 1996). These forecasts are needed for the human brain's capacity coaching decisions. Aside from the high regularity of call arrival series and the intricate variations, including multiple peaks, call arrival data frequently contains many anomalies caused by holidays, special events, promotional activities, and equipment failure. This paper describes a method for forecasting intraday call arrivals that use artificial neural networks (ANNs). In doing so, a quantitative assessment of alternative methodologies for modelling and forecasting anomalies in high-frequency data is performed over several time frames, addressing knowledge gaps of practical significance given the difficulty and cost related to manual inquiry and treatment of such data. It is then preceded by an evaluation of the performance of various ANN modelling methodologies in terms of classification reliability with which normal and exurban durations are modelled. The findings show that ANNs outperform traditional baselines and can model high-frequency anomalies using comparatively straightforward outlier modelling approaches.

Neural networks have many distinct benefits for contact centre forecasting because they do not require complex algorithms to be programmed into them, and they understand from the data provided and can accept external inputs. A neural network can process massive amounts of data from a variety of inputs. The more data a network receives, the better it will identify patterns and trends in that data. It can then make accurate predictions for the future.

The study Barrow and Kourentzes (2018) evaluated the use of various univariate time series forecasting methods for forecasting intraday call arrivals in the presence of outliers. Forecasting ultrasonic call arrivals accumulated in hourly or narrower time buckets remains a significant challenge for call centres. Aside from the complicated intraday and intra-week peaks and Intra year seasonal cycles, call arrival time series analysis includes a vast number of outlier days caused by the incidence of holiday breaks, special events, promotional activities, and system malfunctions. The use of a wide range of univariate time series forecasting techniques for forecasting intraday call arrivals in the presence of such anomalies is evaluated in this study. In addition to traditional statistical methods, artificial neural networks (ANNs) were considered.

The authors affirmed that a variety of methods for encoding the outlying period used include seasonal Moving Average: to identify the order, they measured the SSE over the validation set and chose the order with the least error, exponential smoothing- Combines all characteristics of the time series such as trend and seasonality and then combines whether (none, additive, multiplicative) and choosing the one with the least AIC and Artificial Neural Network - Identifying the appropriate input variables, selecting the hidden node, and configuring the training setup were critical. This is done using regression diagnostics to determine the AR inputs for the ANN. The study of intraday arrival series provides new insights on the impact of outliers on the performance of established forecasting methods. The work has considered outlying periods compared to previous work done. ANNs are the most accurate when only looking into outlying periods; double seasonal exponential smoothing and seasonal weekly moving average perform better in the rest of the periods.

2.4 Hybrid Models

Combining several techniques is an instinctual way to improve prediction performance as the limitations of one method are compensated by the strength of the other model.

The hybrid model's motivation stems from the following aspects. First, determining whether a time series under investigation is created by a linear or non-linear underlying process and whether one technique is more selective than the other in out-of-sample forecasting is frequently tricky in practice (Lin et al., 2017). As a result, forecasters face difficulties selecting the best technique for their specific situations. Second, real-world time series are infrequently either completely linear or completely non-linear. They frequently include both linear and non-linear trends. This means that neither ARIMA nor NNs can be used to model and forecast time series because the ARIMA model cannot handle non-linear relationships, and the neural network cannot model both linear and non-linear trends. As a result, by combining ARIMA and ANN models, complex autocorrelation structures in data can be more precisely modelled (Zhang, 2003). Third, the forecasting literature almost ubiquitously agrees that no solitary method is the best in every scenario. This is primarily because real-world problems are frequently complex, and no single model may be capable of capturing diverse patterns equitably well.

The hybrid system in the study is divided into two steps. The linear part of the problem is analysed using an ARIMA model in the first step, while the residuals from the ARIMA model are modelled using a neural network model in the second step. Because the ARIMA model cannot capture the model's non-linear representation, the linear model's residuals will contain nonlinearity information. The neural network outcomes can predict the error terms for the ARIMA model (Lin et al., 2017). In determining different patterns, the hybrid model takes advantage of the unique features and strengths of both the ARIMA and ANN models.

In the study by Barrow (2016), the authors note that it is tough to conclude if a time series is purely linear or non-linear. Thus, the hybrid approach is best because call centre data is affected by external stimuli such as social changes. The authors improve the Zhang (2003)'s methodology by adding decomposition of unique call centre data into the hybrid model. They note that decomposing a time series model shows its functional characteristics for time series models. They combine the seasonal moving average model with the ANN by looking into the core properties of the intraday data. The authors indicate that by looking into these properties of the data, the model's forecasting accuracy will be increased. Thus the model will be improved by decomposition and combination. The results indicated that the order of SMA would not always have the best accuracy in the hybrid model in different horizons. This means that for short term forecasts, one order would have the best accuracy in SMA, but other orders would outperform the same order in the hybrid model. The authors compare the SMA, SARIMA, Hybrid and Holt winters models on the benchmark comparison. In one bank series, the Hybrid and HW models performed best short term, but the hybrid performed best long term. In another bank series, the hybrid outperformed all models throughout all horizons. Overall the hybrid model performs well. The authors note that using the decomposition of the

SMA models eases the modelling of the hybrid model and further affirms Taylor (2008) that combining the advantages of simple and complex approaches through forecast combination and hybrid methodologies based on time series decomposition is expected to yield benefits.

2.5 Conclusion

To sum up, the forecasting tools include parametric models, non-parametric models and hybrid models. The parametric models include the Autoregressive Moving Average (ARIMA) and Holt-Winters models. These models have been used as benchmark models in time series. They have been noted to perform very well short term. However, the models have fundamental assumptions that must be fulfilled to model the data. The assumptions of stationarity, linearity and normality are restrictive to call centre data affected by external stimuli. Thus, data transformation may be needed to ensure the data fit the assumptions.

The non-parametric models used have shown the ability to perform well when forecasting long term but perform poorly short term. This disadvantages the modelling, as both short-term and long-term forecasts are crucial in running a call centre.

A hybrid of the non-parametric and parametric model would tap into both models' unique features and strength which is an instinctual way to improve prediction performance as the limitations of one model are compensated by the strength of the other model. Since it is hard to confirm whether a series is purely linear or non-linear, the hybrid model from parametric and non-parametric models will have better accuracy. Therefore, from the results from Barrow (2016) where the hybrid model performed well, this study proposes the comparison of the ARIMA and ANN models with their hybrid using the methodology from (Zhang, 2003).

The SMA – ANN hybrid model has been used in bank call centre data from the UK, US and Israel. The banks were open for 18 hrs or less and would operate between 5 to 7 days a week. The question that remains to be investigated based on the present literature is the performance of the ARIMA- ANN hybrid model on a child abuse helpline in a developing country that operates 24 hours a day, seven days a week. The child abuse helpline has unique characteristics which are yet to be investigated.

This study aims to heed to a call by Gans et al. (2003) who detailed that there is little research in call centre forecasting by adding more knowledge on the use of the ARIMA-ANN hybrid model in a Child abuse call centre in a developing country. This methodology has not yet been done in intraday call centre data in a developing country to the best of this study's research.

Chapter 3

Methods

3.1 Introduction

From the conclusions in the literature review, we can see that traditional linear models have a great short term accuracy and non-parametric models have a better long term accuracy when forecasting. A combination of the non-parametric and parametric model should improve the short term accuracy without weakening long term accuracy.

3.2 Data Description

3.2.1 Data Management

This study utilized secondary data from publicly available information. Zindi Africa makes data from Childline Kenya publicly available. Zindi Africa, the custodians of the data, hosts the largest network of African data scientists who use machine learning and AI to solve the world's most critical concerns by bringing together data scientists and organizations (ZindiAfrica, 2022).

The data contains 135,989 records and the variables present are:

1. Calldate- The date and time of the call
2. maincat - Main category call falls into
3. subcat1 - Subcategory call falls under
4. casepriority - Priority of the case

5. referal - Place case referred to
6. caller_gender - Gender of the caller
7. caller_age - Age of the caller
8. caller_county - Area where the call came from
9. child_age - Age of the child in case
10. child_gender - Gender of the child in case
11. child_county - Area where the child is from
12. parent_age - Age of the parent
13. parent_gender - Gender of the parent
14. parent_county - Area where the child is from
15. Abuser_Relationship - Relationship abuser has with the child in case
16. Neglector_Relationship - Relationship neglector has with the child in case
17. Physical_abuser_Relationship - Relationship physical abuser has with the child in case

The variable calldate was used to create a univariate time series data, grouping the total number of calls received in an hour. The data available was collected between 1^{st}January 2016 to 12^{th} July 2016.

3.3 Data Analysis

The childline call frequency data is time-series data as its data points are indexed in time order which is the recorded time of the call. The time series may have underlying trends, cycles and irregular fluctuations. Different models express these variations uniquely while forecasting.

From the literature review, it was prudent to note how traditional parametric models such as ARIMA and non-parametric Models such as Neural Networks model the linear and non-linear components of the call centre data. Therefore, the models ARIMA, ANN and Hybrid ARIMA-ANN were used in the analysis.

3.3.1 ARIMA

The Auto-Regressive Integrated Moving Average Model is used to model non-stationary Time series data that has been transformed to stationary through differencing (Cryer and Chan, 2008). y_t is an ARIMA model iff the d^{th} difference denoted by

$$z_t = \nabla^d y_t \tag{3.1}$$

is a stationary ARMA model. Thus if z_t is ARMA(p,q) then y_t is ARIMA (p,d,q)

If y_t is a non stationary time series modelled as ARIMA(p,d,q) then:

$$y_t = \nabla^d y_t \sim ARMA(p,q) \tag{3.2}$$

with z having a constant mean of $\mu = 0$ because of the stationarity.

AutoRegressive Processes

The Auto-Regressive model undertakes regression on itself. A p^{th} order regressive process AR(p) y_t is given by:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + w_t \tag{3.3}$$

The current values of the series y_t is a linear combination of the p most recent past values of itself and an innovation term w_t which incorporates everything new in a time series not captured by the past values (Cryer and Chan, 2008).

Moving Average Processes

The Model uses past errors as the explanatory variables regression on itself. A q^{th} order Moving average process MA(q) y_t is given by:

$$y_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} \tag{3.4}$$

with $w_t(1,2,3)$ being a white noise process of iid random variables.

$$w_t \sim N(0, \sigma^2) \tag{3.5}$$

(Cryer and Chan, 2008).

Because the model is described in terms of past errors, we estimate the coefficients $\theta; (\theta = 1, 2, \ldots)$ and then use the forecasting model. Only q errors will affect the current level y_t, but higher-order errors will not.

ARMA

Mixed Auto-Regressive Moving Average model is part Auto-Regressive and part Moving average (Cryer and Chan, 2008). The model assumes that the series is stationary. The ARMA of order p & q, i.e. ARMA(p,q) equation, is of the form:

$$y_t = \sum_{i=1}^{q} \theta_i w_{t-i} + \sum_{j=1}^{p} \phi_j y_{t-j} + w_t \tag{3.6}$$

Time series data is very dynamic thus is usually not stationary and may be of the form $Y_t = \mu_t + w_t$ with μ_t being a non constant mean function and w_t being the white noise.Differencing of the series transforms the model to stationary. Differencing is the change between consecutive observations in the original series such that the first difference is:

$$\triangledown Y_t = Y_t - Y_{t-1} \tag{3.7}$$

and the second difference:

$$\bigtriangledown^2 Y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \tag{3.8}$$

leading to a general type of model.

The ARIMA (p,d,q) would now be:

p - indicating the number of Autoregressive Terms,

q - Number of Moving average Terms

d - Number of times differenced

The Box Jenkins Methodology as stated in Box et al. (2015) shall be followed. The methodology has the following steps:

1. Phase 1: Identification - This establishes the time series's stationarity. Transformation to stabilize the variance, Differencing to obtain a stationary series, and examining the ACF and PACF to identify potential models.

2. Phase 2: Estimation and Testing - Choosing the parameters in the model and use of AIC to select the best model. This phase is iterative with Phase 1 until the best model is arrived at, where the residuals are normally distributed.

3. Phase 3: Application - Using the Model to forecast.

Seasonality Seasonal AR(p) model of order P and seasonal period s is defined by :

$$y_t = \phi_1 y_{t-s} + \phi_2 y_{t-2s} + \ldots + \phi_p y_{t-ps} + w_t \tag{3.9}$$

as shown by (Cryer and Chan, 2008).

The q^{th} order of a Seasonal Moving average process MA(q) y_t is given by:

$$y_t = w_t + \theta_1 w_{t-s} + \theta_2 w_{t-2s} + \ldots + \theta_q w_{t-qs} \tag{3.10}$$

The Seaonal ARMA is denoted as : ARMA(p,q)×(P,Q)$_s$. In case of a non stationary seasonal time series model, seasonal differencing is done Cryer and Chan (2008), it is denoted by $\bigtriangledown_s Y_t$. as defined as:

$$\bigtriangledown_s Y_t = Y_t - Y_{t-s} \tag{3.11}$$

The Seasonal ARIMA will now be defined as: ARIMA(p,d,q)×(P,D,Q)$_s$. Many periods in a ARIMA model are denoted by: ARIMA $(p,d,q)(P_1,D_1,Q_1)_{s_1}(P_2,D_2,Q_2)_{s_2}$ (Mohamed et al., 2010). y_t is an SARIMA model iff the D^{th} difference denoted by

$$z_t = \bigtriangledown^d \bigtriangledown_s^D y_t \tag{3.12}$$

is a stationary SARMA model. Thus if z_t is SARMA(p,q) (P,Q)$_s$ then y_t is SARIMA (p,d,q) (P,D,Q)$_s$

The stationarity tests were conducted using the KPSS test (Kwiatkowski Phillips Schmidt Shin) and ADF.test. The KPSS test uses linear regression to break down a time series into a deterministic trend (βt), random walk (r_t) and stationary error (ε_t) as:

$$x_t = r_t + \beta t + \varepsilon_t \tag{3.13}$$

The test checks for non-stationarity due to a unit root or stationarity about the linear trend or mean. The Augmented Dickey-Fuller test also tests unit-roots that indicate non-stationarity.ADF test expands the Dickey-Fuller test equation to include high order regressive process in the model.

The Autocorrelation plot was used to obtain the order of the MA process, while the Partial Autocorrelation plot was used to obtain the order of the AR process. In the testing phase, it was critical to evaluate the residuals and the ACF of the residuals. The p-value of the Ljung-Box Q statistic was used to test the residuals. The Ljung-Box Q statistic tests if the white noise are iid, indicating if autocorrelations from the errors are zero. The test statistic can be calculated as :

$$Q = n(n+2) + \sum_{k=1}^{K} r_k^2 (n-k)^{-1} \tag{3.14}$$

where K= time lag and r_k =the accumulated sample autocorrelations (Box et al., 2015).

The final stage was to forecast with the estimated model.

3.3.2 Neural Networks

Time series data is not always stationary, linear or normal. While transforming the data to stationary, critical information may be lost. Non Parametric approaches such as Neural Networks and Spectral Analysis do not require these assumptions.

Artificial neural networks (ANNs) are forecasting techniques based on simple brain mathematical models. Using them, complex non-linear relationships between the response variable and its predictors are possible. They hold several advantages over traditional Time series models as they support missing values.

Neural Network Architecture

A neural network consists of layers of neurons. The last layer would be the predictor or inputs, with the first layer being the forecast. In linear regression, which is a simple network, we have two layers, i.e. the predictor and the output, but in Neural networks, there are intermediary layers between the inputs and the outputs, making it non-linear.

In Multi-layer feed forward networks each layer receives input from previous layer as an output and is then the input to the next layer. The relationship between the output y_t with the input $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ can be represented as:

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g \left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i} \right) + \varepsilon_t \tag{3.15}$$

where $\alpha_j(j = 0, 1, 2, \dots, q)$ and $\beta_{ij}(i = 0, 1, 2, \dots, p;\ j = 0, 1, 2, \dots, q)$ are the model parameters which are weighted linear combinations called connection weights. p is the number of input nodes and q as the number of hidden nodes.

The hidden layer g is modified using a non-linear function.

$$g(x) = \frac{1}{1 + e^{-}x} \tag{3.16}$$

which acts by lessening the impact of extreme input values, making the network more resistant to outliers.Therefore model 3.15 maps nonlinear past values $(y_{t-1}, y_{t-2}, \dots, y_{t-p})$ to future value y_t:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, w) + \varepsilon_t \tag{3.17}$$

with w as a vector of all parameters and f is the function of the network and connection weights. This equation shows how the neural network is a non-linear autoregressive model. The lagged values of the time series are the inputs in a neural network.

NNAR

NNAR(p,q) indicates a Neural Network Autoregression model with p lagged inputs and q nodes in the hidden layer. Seasonal NNAR is indicated by NNAR(p,P,q)$_m$ with P as the number of lagged inputs at lag m .The hidden layer choice, which is data-dependent, is essential to avoid overfitting. The choice of p helps determine the autocorrelation structure of the time series, which is non-linear. Parameter estimation is done by choosing the ones that reduce the accuracy metrics, and the model is evaluated using the out of sample prediction. The network forecasts iteratively one step at a time; thus, for more than one step forecast, the first forecast is used as the following input along with the historical data until all forecasts have been computed.

3.3.3 Hybrid - ARIMA & ANN

Using the hybrid model, a time series can have both linear and non-linear components. The additive model is as shown below.

$$y_t = L_t + N_t \tag{3.18}$$

L_t is the Linear component, and N_t is the non-linear component. The ARIMA will model the linear component first with its residuals having the non-linear relationship Zhang (2003) which can be written as:

$$e_t = y_t - \hat{L} \tag{3.19}$$

With $\hat{L}$ being the forecast from the ARIMA model.

Residual analysis indicates the linear patterns of residuals, thus confirming if the linear model is adequate or if there are still linear correlations in the model. Non-linear residual patterns do not have a clear diagnostic check; thus, the patterns may be revealed by modelling the residuals using non-linear models such as ANNs. The residual ANN model is:

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + \varepsilon_t \tag{3.20}$$

ε_t is the random error, and f is the neural network's non-linear function as shown by (Zhang, 2003). The forecasts from both models would be:

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \tag{3.21}$$

3.4 Forecasting and Forecast Evaluation

The analysis was conducted on the total call frequency per hour. The validation set technique, which involves splitting the data into the test and training set, was used. Where the training set was used to obtain the model parameters, and the test set was used to confirm the accuracy level of the data set. The next step waw to predict future values.

To test the performance of the three procedures, Mean Absolute Error(MAE) was used. MAE was used because the Root Mean Squared Error (RMSE) is sensitive to outliers and the data has small counts of 1 which is not ideal when using Mean Absolute Percentage Error (MAPE) (Hyndman and Koehler, 2006).

For the out of sample prediction, the model's accuracy was be judged by the error statistic:

$$MAE = mean(|e_t|) \tag{3.22}$$

where$e_t = (Y_t - F_t)$, Y_t is the Observation at time t and F_t is the Forecast of Y_t. The lower the MAE, the better.

To further determine if the forecasts from the models were significantly different the Diebold-Mariano test was conducted.The Diebold-Mariano test compares the forecast accuracy of two forecast methods (Diebold and Mariano, 2002). The test statistic is as:

$$\frac{\bar{d}}{\sqrt{[\gamma_0 + 2\sum_{k=1}^{h-1} \gamma_k]/n}} \tag{3.23}$$

The time series d_i is called the loss-differential which is the difference between the residuals of the forecasts of the models being compared and $\hat{d}$ is the mean of the differences for all observations i, γ_k is the autocovariance at lag k and $h = n^{\frac{1}{3}} + 1$.

The modelling was performed using the software R studio. The packages zoo was used to impute the missing timelines and convert the data into time series data. The smooth package was used in the modelling ARIMA models and the forecast package was used in the modelling ANN model and the forecasting from all the models.

Chapter 4

Results

4.1 Exploratory Data Analysis

When we acquire a dataset, a very helpful initial step is to carefully examine it to determine what it contains.This can be done using exploratory data analysis(EDA).EDA is useful to ascertain and improve the accuracy of our data and our forecasts (Pearson, 2018). In this section we explored graphical characterization and decomposition of the time series data to understand the data.

4.1.1 Data

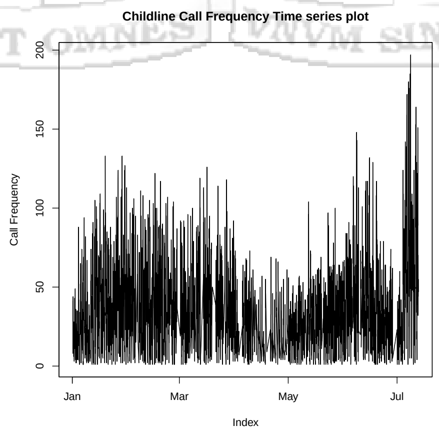
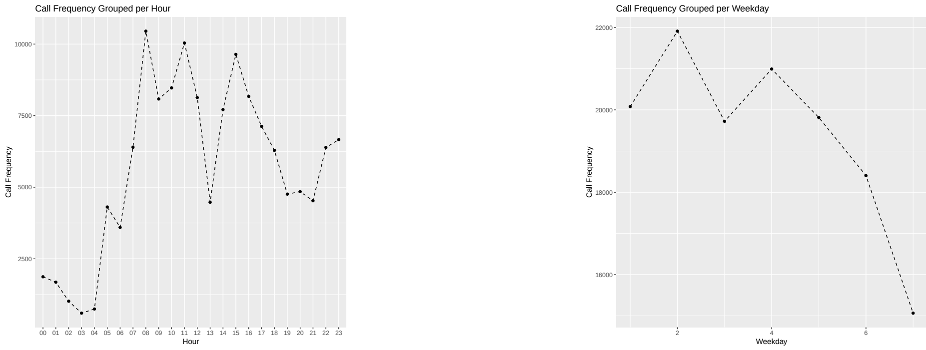


Figure 4.1: Run Sequential Plot of the Childline Data

The data used was collected from Childline Kenya from 1st January 2016 at 7 am to 12th July 2016 at 11 pm. The data contains 135,988 records of individual calls recorded within the time range. The calls were then grouped by the hour to create hourly totals per day, resulting in 4,649 observations after the imputation of missing hours. The missing hours were



(a) Total Calls Received within a day   (b) Total Calls received within a Week

Figure 4.2: Line Plots per Day and per Weekday

imputed using the function na.approx() in R Studio. The function uses linear approximation to interpolate the NAs due to missing hours. The method was considered appropriate because of the high frequency nature of the data.

The run sequential plot figure 4.1 reveals a decrease in calls in April but a steady increase from May to July.

Line plots shown in figure 4.2 were created to reveal periodicity within a day and a week. Figure 4.2(a) indicated peaks at 8 am, and 11 am with a decline at 1 pm and another peak at hour 15. Figure 4.2(b) indicated a decline of calls during the weekend with a peak on Tuesdays.

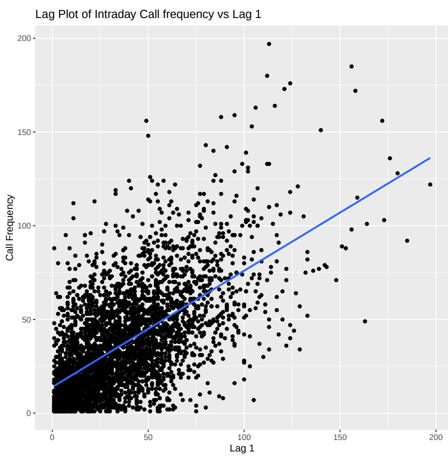
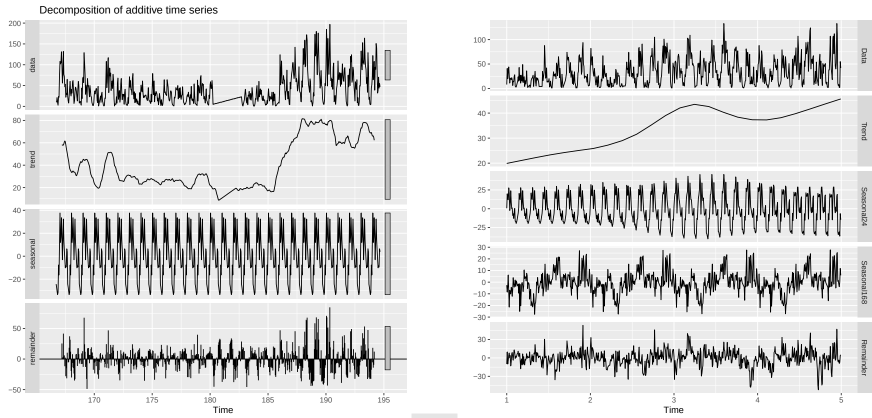


Figure 4.3: Lag plot at lag 1

A lag plot at lag 1, figure 4.3, was used to assess the randomness of the intraday data.



(a) Decomposition of the Hourly Call Volume data

(b) Decomposition with Multiple seasonal Periods

Figure 4.4: Decomposition of the Time series data

```
        Augmented Dickey-Fuller Test

data:  ts.diff.day
Dickey-Fuller = -15.403, Lag order = 16, p-value = 0.01
alternative hypothesis: stationary
```

(a) ADF Test

```
        KPSS Test for Trend Stationarity

data:  ts.diff.day
KPSS Trend = 0.0068009, Truncation lag parameter = 10, p-value =
0.1
```

(b) KPSS Test

Figure 4.5: Stationarity Tests

Decomposition of a time series is required to estimate the components of a time series: trend, seasonality and irregular components (Coghlan, 2015). Figure 4.4(a) was used to plot the Childline data with only one periodicity component at a frequency of 24 as daily data. Then figure 4.4(b) was used to plot the Childline data with periods 24 and 168 indicating daily and weekly periods.

4.1.2 Data Transformation

The data is transformed using differencing to eliminate the periodicity as per the steps in the Box-Jenkins methodology. Differencing is done to remove weekly periodicity and then another differencing is done to remove daily periodicity.

Stationarity tests were conducted after transformation as shown in figure 4.5. The Dickey Fuller tests for serial autocorrelation with the hypothesis as:

$$H_0 : \textit{There is a unit root}$$

$$vs$$

$$H_1 : \textit{Time series is stationary}$$

Cryer and Chan (2008) Figure 4.5(a) shows that the p-value is < 0.05. The KPSS stationarity determines if a time series is stationary around a mean or linear trend, or non-stationary because of a unit root (Kokoszka and Young, 2016).

$$H_0 : \textit{Data is Stationary}$$

$$vs$$

$$H_1 : \textit{Data is not stationary}$$

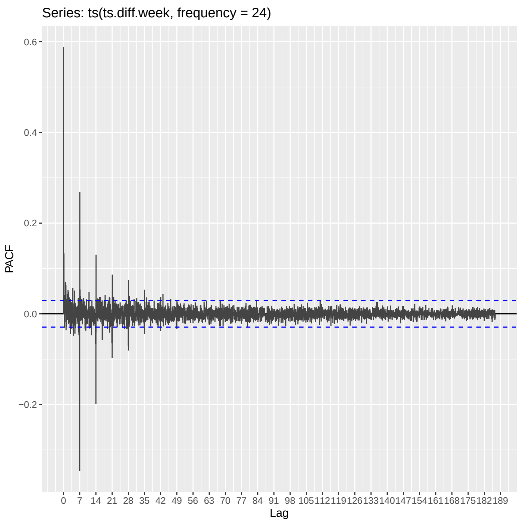
Figure 4.5(b) shows that the p-value is > 0.05.

4.2 Models

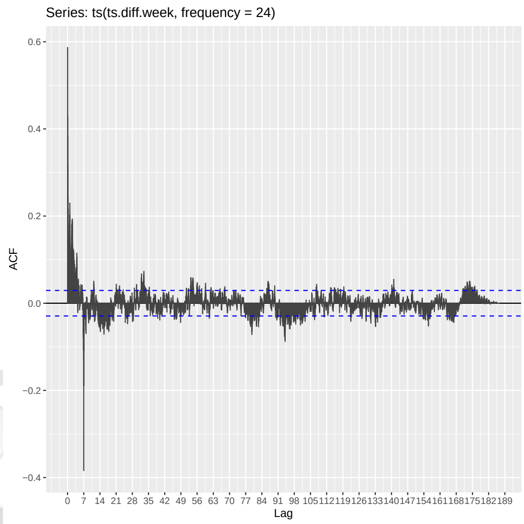
4.2.1 ARIMA

Using the Box-Jenkins methodology, the identification stage involved plotting PACF and ACF of the data after being differenced daily and weekly. The figures 4.6 shows the orders after the weekly and daily differencing.

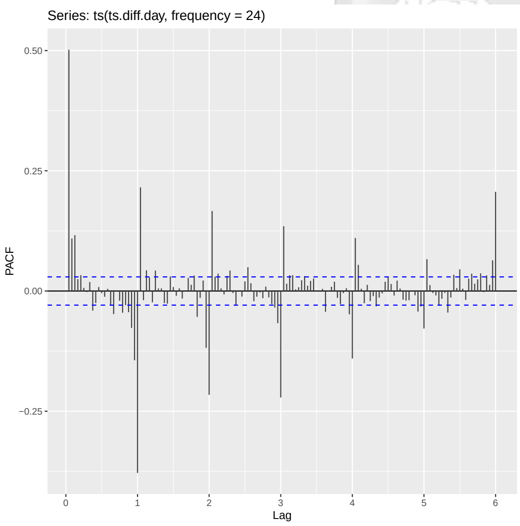
From the orders in the figure 4.6 different possible ARIMA models were created in order to choose the best model.Training of the model was done by choosing the model with the least AIC. The Table 4.1, shows the AICs of the different possible models.



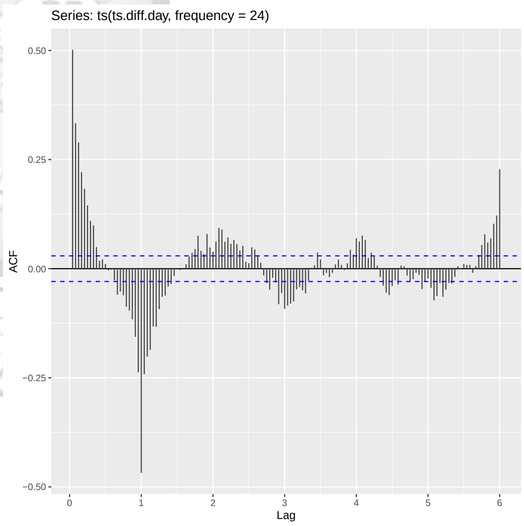
(a) PACF of the weekly differenced data



(b) ACF of the weekly differenced data



(c) PACF of the daily differenced data

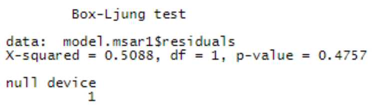


(d) ACF of the daily differenced data

Figure 4.6: ACF and PACF of the transformed data

Table 4.1: IC(s) of each possible ARIMA model

	AIC	AICc	BIC	BICc
SARIMA(1,0,1)[1](0,0,1)[24](5,0,1)[168]	**31118**	**31118**	**31180**	**31180**
SARIMA(1,0,1)[1](0,0,1)[24](5,0,0)[168]	31416	31416	31472	31472
SARIMA(1,0,1)[1](0,0,1)[24](4,0,1)[168]	31120	31120	31176	31177
SARIMA(1,0,1)[1](0,0,1)[24](4,0,0)[168]	31545	31545	31595	31595
ARIMA(4,1,2)	32304	32305	32348	-

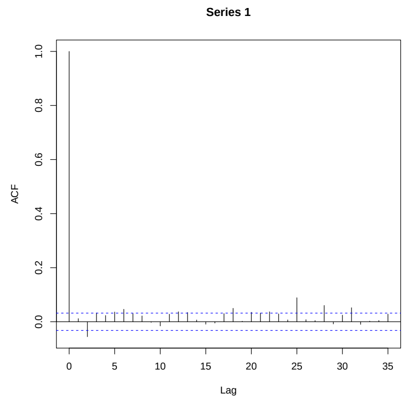


```
        Box-Ljung test

data:  model.msar1$residuals
X-squared = 0.5088, df = 1, p-value = 0.4757

null device
          1
```

(a) Box Test



(b) ACF of the Residuals

Figure 4.7: Residual Analysis

Residual analysis is then conducted as a diagnostic tool to ensure no lack of fit. As stated by Box et al. (2015) a visual check of the residuals should be the first step. The ACF plot figure 4.7(b) was used as the first visual check of the residuals. The Ljung-Box Test checks that the first 10-20 autocorrelations of the model imply that the model is insufficient (Box et al., 2015). The hypothesis is:

$$H_0 : \textit{Data does not show lack of fit}$$

$$vs$$

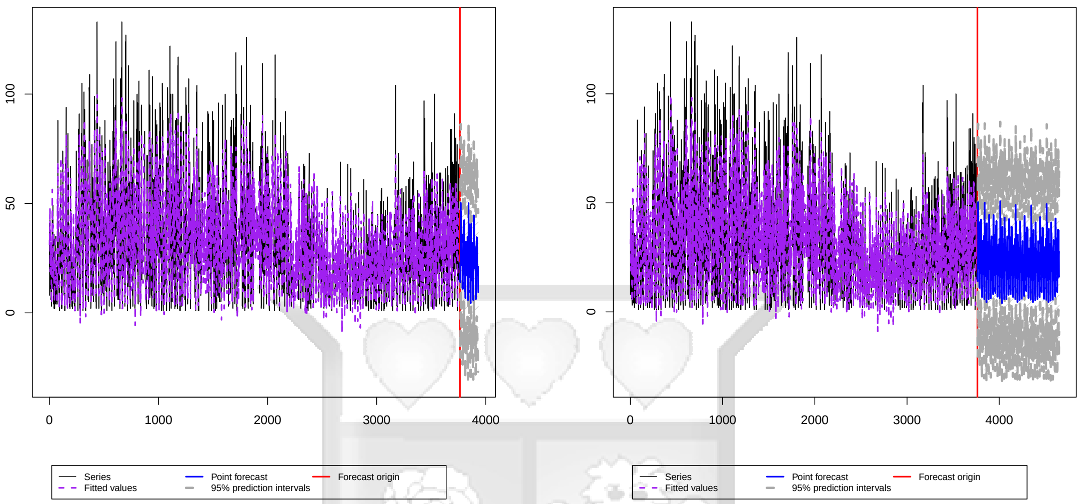
$$H_1 : \textit{Data dies show lack of fit}$$

The figure 4.7(a) shows that the p-value is > 0.05.

The final stage in the Box-Jenkins is forecasting. Short term and long term forecasts were chosen for periods of 1 week and 36 days. Figure 4.8 shows the results of the ARIMA forecast compared to the test data values.

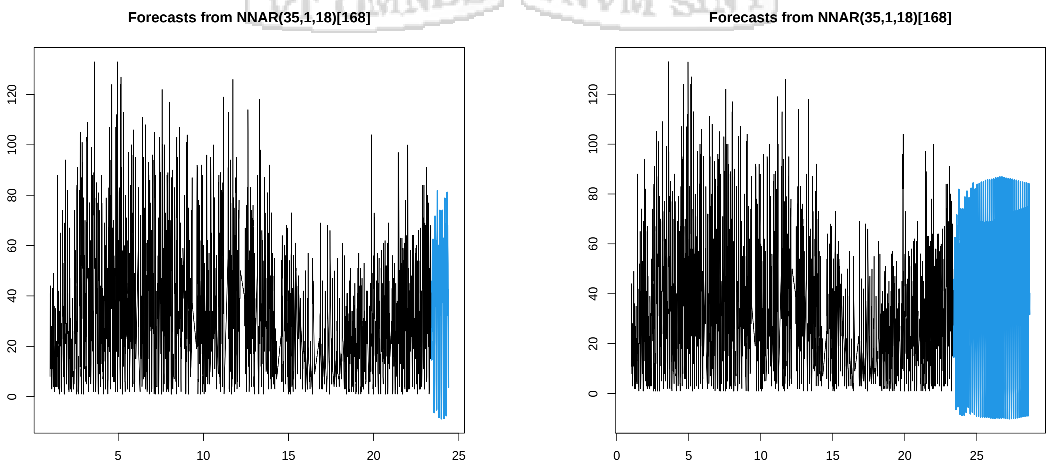
4.2.2 NNAR

The function nnetar() selectively chooses the most appropriate model for the data based on forecast results. The model NNAR(35,1,18)[168] was used on the Childline data.The model



(a) ARIMA Forecast for short term period (b) ARIMA Forecast for long term period

Figure 4.8: ARIMA Forecasts



(a) NNAR Forecast for short term period (b) NNAR Forecast for long term period

Figure 4.9: NNAR Forecasts

Table 4.2: Training the Hybrid-ANN parameters

	ME	RMSE	MAE	MPE	MAPE
NNAR(2,5,1)[168] repeats=20	1.26	13.38	10.26	128.43	131.91
NNAR(2,5,1)[168] repeats=40	**1.21**	**13.36**	**10.25**	**124.00**	**128.59**
NNAR(2,5,1)[168] repeats=30	1.23	13.37	10.26	122.82	128.42
NNAR(1,5,1)[168] repeats=30	1.55	13.45	10.30	116.35	123.85
NNAR(3,5,1)[168] repeats=40	0.91	13.37	10.27	139.90	151.44
NNAR(2,6,1)[168] repeats=40	1.37	13.39	10.24	124.87	149.14
NNAR(2,4,1)[168] repeats=40	1.27	13.41	10.26	118.93	140.68
NNAR(2,5,2)[168] repeats=40	1.31	13.40	10.27	141.57	145.74
NNAR(2,5,3)[168] repeats=40	1.47	13.43	10.28	142.50	146.78

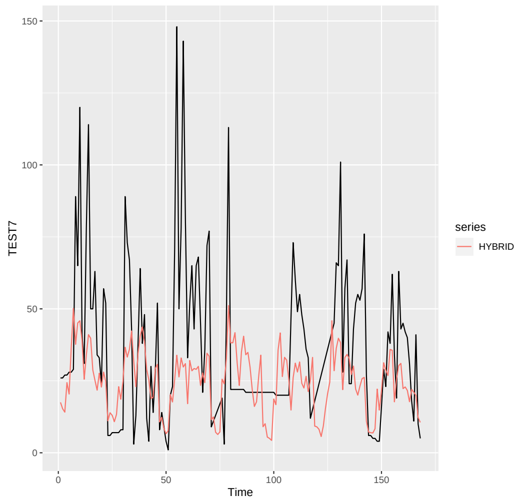
is of the form NNAR(p,P,q)$_m$ with 168 as the period m, 18 as the hidden node q , 35 as the lagged inputs and 1 as the lagged inputs P at period m. The Figure 4.9 shows the results of the NNAR forecast compared to the test data values.

4.2.3 Hybrid: ARIMA-NNAR

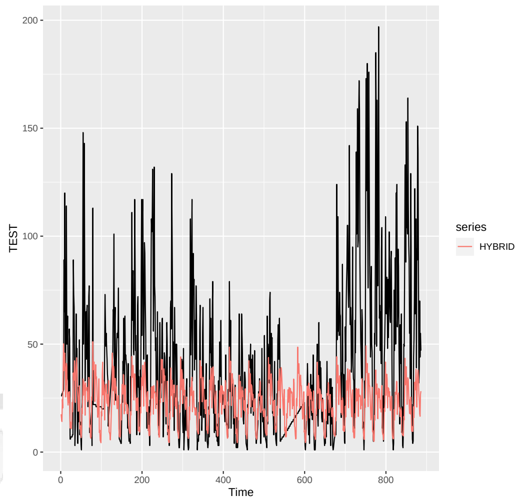
The best ARIMA model was used in the hybrid methodology. Training of the model was done by choosing the NNAR (p, P,q)$_m$ model parameters with the least RMSE. In this study, we selected repeats between 20-40 at random. The choice of P was based on the number of seasonal lags in the ARIMA model, with five as the selected start. The choice of p was also determined by the number of lags for the short term autocorrelation in the model. The choice of the hidden layer was also chosen at random from one single hidden layer. Table 4.2 shows the prospective models included while training the model. NNAR(2,5,1)[168] with 40 repeats had the least RMSE out of sample.

The model was then used on the residuals from the SARIMA model, and the forecasts from both models were added to create $\hat{y}$. $\hat{y}$ was then tested against the Childline data test set as the out of sample prediction.

Figure 4.10 shows short term and long term forecasts by the hybrid model.



(a) Hybrid Forecast for short term period



(b) Hybrid Forecast for long term period

Figure 4.10: Hybrid Forecasts

Table 4.3: 7 days forecast:Accuracy Metrics of the Models

	MAE
SARIMA	18.17
NNAR	19.88
Hybrid	18.11

4.3 Comparative Forecasting Accuracy Evaluation

The accuracy metric used in the study was MAE. MAE is a is a scale dependent accuracy measure useful when evaluating a forecasting method's ability to predict the series' median future values (Koutsandreas et al., 2021). Table 4.3 shows the metrics of short term forecasts whereas Table 4.4 shows the metrics of long term forecasts.

Further, the out of sample MAE was calculated for each day the figure 4.11 shows the gradual changes of MAE from day 1 to day 36.

Table 4.4: 36 days forecast: Metrics of the Models

	MAE
SARIMA	22.70
NNAR	24.16
Hybrid	22.60

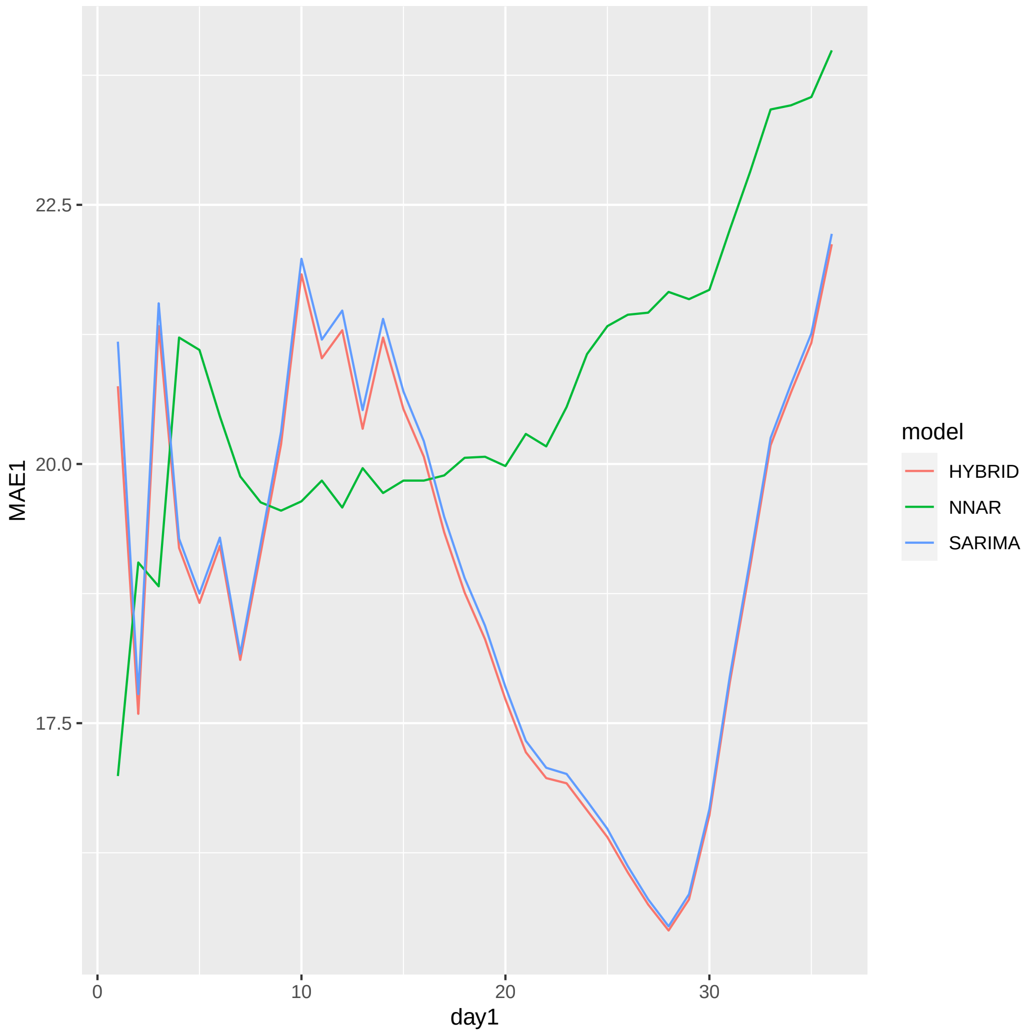


Figure 4.11: MAE of the Childline Data from One day to 36 Days

We utilized the Diebold and Mariano test statistics to determine how much the hybrid model forecast differed from the constituent models forecast for forecasts with a horizon of h = 886 on the test data. The Diebold and Mariano static checks if there is a difference between two forecasts (Diebold and Mariano, 2002). The hypothesis is:

$$H_0 : \textit{No difference in the accuracy}$$

$$\textit{vs}$$

$$H_1 : \textit{Difference in the accuracy}$$

The p-value between the ANN and Hybrid forecasts was 0.001 whereas the p-value between the ARIMA and Hybrid forecasts was 1.

Chapter 5

Discussion

5.1 Data

The data was analysed as a univariate time series model with a frequency of 24 as it is daily data. The run sequential plot figure 4.1, showed that the data was not stationary; there was no consistent trend because the call volume fluctuated. The plot indicates possible annual seasonality when the call volumes decrease in April and increase in July. However, it is impossible to estimate the annual seasonality because the available data is collected for only seven months.

The daily line plot figure 4.2(a), shows a decrease in calls from 12 am to 3 am with a steady increase from 5 am and the highest peak at 8 am. The fluctuations indicate periodicity as there are peak hours of the calls implying that many respondents would wait until 8 am to seek help through the childline. It would be prudent for the childline to sensitise the community that the helpline is a 24-hour service. There is also a decline at 1 pm with another increase at 3 pm. The weekly plot figure 4.2(a), also indicated periodicity as there is an apparent decrease in calls during the weekend. The lag plot shows a mild positive relationship at lag 1, indicating a potential AR model.

In the decomposition on the series figure 4.4(a), indicated an hourly periodicity with an irregular trend. This trend pattern could be due to an untapped additional seasonality from a more extended frequency, in this case, rendering it multi-seasonal data. Therefore further decomposition was done based on weekly periods of 168. The trend became apparent when the time series was plotted as multi-seasonal data, as in figure 4.4(b). Both weekly and daily periods are apparent in the figures.

5.2 Data Transformation

The data is not stationary from the time series plot figure4.1 and the decomposition in figure 4.4; this may be due to the irregular trend and multiple periodicities present. The ARIMA model and ARIMA constituent in the hybrid require stationarity as mandated in the Box-Jenkins methodology. The data is transformed using differencing to eliminate the weekly and daily periodicity. After the transformations, the Dickey-Fuller test shows stationarity because the p-value is < 0.05. The KPSS stationarity test also affirms stationarity because the p-value is > 0.05; thus, we fail to reject the null hypothesis.

5.3 Modelling

The next step in the Box-Jenkins methodology is to identify the orders. The ACF and PACF of the weekly differenced and daily differenced data are created. First, we put the aperiodic short-term correlation part of the model in place. The ACF diagram and PACF diagram showed the sequence after 24 steps and 168 steps of difference. According to the results, Figure4.6 showed that the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) were not truncated within the 24^{th} order. Thus to express short-term autocorrelation information, we utilised a low-order ARMA (1,1).

We then observed the autocorrelation coefficient and partial autocorrelation coefficient with the cycle length as the unit of delay in the 24^{th}, 48^{th} order, ... for the autocorrelation characteristics of the daily cycle. Except for the delay 24^{th} order, the autocorrelation coefficients were not 0 as in figure 4.6(d).In Figure 4.6(c), there were many significant partial autocorrelation coefficients with a truncation feature, so for the 24-step cycle, ARIMA(0,1)[24] is used to express the autocorrelation.

To judge the autocorrelation coefficient and partial autocorrelation coefficient for the weekly cycle, we plotted ACF and PACF in order of 168, thus 168^{th}, 336^{th}, ... From figure 4.6(b), the ACF is significant at the 168^{th} order. Figure 4.6(a) showed that the PACF is significant within order 840. Thus ARIMA(5,1)[168] was used to express the autocorrelation of the weekly cycle.The model is in the form SARIMA $(p,d,q)(P_1,D_1,Q_1)_{s_1}(P_2,D_2,Q_2)_{s_2}$ with p,d,q being the low-orders, P_1,D_1,Q_1 being the orders of the 24^{th} period s_1 and P_2,D_2,Q_2 being the orders of the 168^{th} period s_2.

The data is then split into a training and testing set using an 80:20 ratio. Five comparative models were used. The ARIMA (4,1,2) was created using the auto.Arima() function models the best ARIMA model based on AIC (Coghlan, 2015). The msarima() function was used to model the Multi-periodicity SARIMA models. The SARIMA(1,0,1)1(0,0,1)24(5,0,1)168 has the least AIC; thus, it was used in the modelling.

Residual analysis was conducted to check the model's goodness of fit. The ACF plot figure 4.7(b), showed that there were not many significant autocorrelations in the residuals, indicating that the model does not show a lack of fit. The Ljung-Box result in figure 4.7(a) affirms this since we fail to reject the null hypothesis at a p-value of 0.4757, showing that the model indeed does not show a lack of fit. The residual analysis results indicated that the model could now be used for the final stage of the Box-Jenkins methodology, which is to forecast.

The hybrid methodology required the Neural network model training as stipulated in the (Zhang, 2003). The training was achieved by using the model parameters that had the best accuracy in the out of sample prediction of the residuals. Cross-validation was done by further splitting the residuals into a training and testing set. Different parameters (p, P, q)$_m$ were used. Table 4.2 shows that an increase of the hidden nodes from a size of 1 increased the RMSE. The increase in RMSE after increasing the hidden layers is in line with results from Triebe et al. (2019) which show that a single hidden layer is sufficient in Neural Network modelling. An increase or a decrease from 5 n P, the seasonal lag, would also increase the RMSE as P=5 is the inherent order from the SARIMA model at [168] periods. The best choice of repeats was 40, with the best choice of p as 2. The NNAR(2,5,1)[168] with repeats of 40 produced the best out of sample results with the least RMSE and MAPE. The model was used on the SARIMA residuals, and its results $\hat{N}_t$ were added to the forecasts from the SARIMA model $\hat{L}_t$ to create y_t.

5.4 Comparative Forecasting Accuracy Evaluation

The short term results in the table 4.3 showed that the hybrid model was the most accurate method short term based on the MAE. This result follows the results shown by (Barrow, 2016).The hybrid model reduced MAE of the NNAR model. The accuracy indicated that the model tapped into the SARIMA model's advantages without compromising the accuracy as hypothesised in this study. Overall the hybrid model outperformed the constituent models in short term forecasts by being the most accurate model.

The long term results in table 4.4 show that the hybrid model had the best performance of the three models. The hybrid model improves the long term ARIMA forecast from MAE 22.70 to 22.60, indicating that the NNAR component in the hybrid model is utilised to better the long term forecast.

We then compared the forecasts to ensure that the results from the accuracy metrics were significant. The results from the Diebold Mariano tests indicate a significant difference between the ANN and Hybrid forecasts with a p-value of 0.01. However, with a p-value of 1, the forecasts were not significantly different between the SARIMA and the Hybrid model. These results are in conjunction with 4.11, which shows a gap between the ANN model with the hybrid model, whereas the ARIMA and hybrid model seemed to have close results.

Overall, the hybrid model performed better than its constituent models based on MAE. In this study, we have shown that the hybrid model enhances the short-term accuracy without compromising long-term performance, indicating that the advantages of each constituent model have been utilised.

Chapter 6

Conclusion and Recommendations

6.1 Conclusion

This study finds, based on the results described in the previous Chapter, the hybrid model consisting of the ARIMA and ANN model significantly improves the ANN model in short term accuracy and marginally improves the ARIMA model in long term accuracy.

These results could be helpful to Childline Kenya and call centres in general because the hybrid model can be used in place of its constituent models. Thus, one model can be used instead of using two different models for different horizons.

6.2 Limitations

The Childline data available is dated 1st January to 12th July; thus, the time-series component of annual seasonality could not have been included in the model. The run sequential plot indicated a decrease in April and an increase in July. This component may have been crucial to creating more accurate results.

6.3 Recommendations

Future work could look into the Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components(TBATS) Model and the Prophet model, which are models useful in complex periodicity. Comparing the ARIMA and TBATS model with the ensemble could be critical in call centre data containing complex periodicity. In addition, the

Prophet model could be explored for Call centre data as the model is very suitable for daily data forecasting. The model includes weekly periodicity, yearly seasonality, and holiday effects. This unique attribute of the prophetic model could prove helpful for Call centre data as the use of linear models would require stationarity; thus, holiday effects are usually removed.

References

Al-Azzani, M. A., Davari, S., and England, T. J. (2020). An empirical investigation of forecasting methods for ambulance calls-a case study. *Health Systems*, pages 1–18.

Barrow, D. and Kourentzes, N. (2018). The impact of special days in call arrivals forecasting: A neural network approach to modelling special days. *European Journal of Operational Research*, 264(3):967–977.

Barrow, D. K. (2016). Forecasting intraday call arrivals using the seasonal moving average method. *Journal of Business Research*, 69(12):6088–6096.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

ChildlineKenya (2012). Child sexual abuse in kenya:occurance, context, risk factors and consequences. pages 3–25.

ChildlineKenya (2022). Home / childline kenya. https://childlinekenya.co.ke/.

Coghlan, A. (2015). A little book of r for time series. *Published under Creative Commons Attribution*, 3.

Cryer, J. D. and Chan, K.-S. (2008). *Time series analysis: with applications in R*, volume 2. Springer.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Fildes, R. and Kumar, V. (2002). Telecommunications demand forecasting—a review. *International journal of forecasting*, 18(4):489–522.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Hill, T., O'Connor, M., and Remus, W. (1996). Neural network models for time series forecasts. *Management science*, 42(7):1082–1092.

Huang, H., Jiang, M., Ding, Z., and Zhou, M. (2019). Forecasting emergency calls with a poisson neural network-based assemble model. *IEEE Access*, 7:18061–18069.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Kokoszka, P. and Young, G. (2016). Kpss test for functional time series. *Statistics*, 50(5):957–973.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., and Assimakopoulos, V. (2021). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, pages 1–18.

Lin, T., Guo, T., and Aberer, K. (2017). Hybrid neural networks over time series for trend forecasting.

Mohamed, N., Ahmad, M., Ismail, Z., and Suhartono, S. (2010). Short term load forecasting using double seasonal arima model. *Statistics Faculty of Computer and Mathematical Sciences*, 15:57–73.

MtotoNews (2022). Mtoto news - children organization. https://www.mtotonews.com/.

Pearson, R. K. (2018). *Exploratory data analysis using R*. Chapman and Hall/CRC.

Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2):253–265.

Taylor, J. W. (2010). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting*, 26(4):627–646.

Triebe, O., Laptev, N., and Rajagopal, R. (2019). Ar-net: A simple auto-regressive neural network for time-series. *arXiv preprint arXiv:1911.12436*.

Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175.

ZindiAfrica (2022). Zindi -learn, grow, build connections and help transform the world with ai. https://zindi.africa/.

Appendix A

R Script

The R code used for EDA and Modelling in Chapter 4.

```
knitr::opts_chunk$set(fig.pos = "H", echo = FALSE, warning=FALSE, fig.align="center
#deleting previous datasets
rm(list=ls())
#Reading in the data and libraries and setting work directory
setwd("C:/Users/grace/OneDrive/Documents/2. RESEARCH METHODS/Thesis")
library(tseries)
library(dplyr)
library(ggplot2)
library(zoo)
library(knitr)
library(kableExtra)
library(tidyverse)
library(car)
library(tidyr)
library(forecast)
library(ggbio)
#Preparing the data
data <- read.csv(file = "childline.csv",header = TRUE,stringsAsFactors = FALSE)
#changing the formats of the variables
childline<- data %>%
mutate(date=substr(data$calldate,1,10))%>%
mutate(year=trimws(substr(calldate,1,4)))%>%
```

```
mutate(month1=trimws(substr(calldate,6,7)))%>%
mutate(day=trimws(substr(calldate,9,10)))%>%
mutate(time=trimws(substr(calldate,12,13)))%>%
mutate(dttime=as.POSIXct(calldate, format="%Y-%m-%d %H"))%>%
mutate(time_index=paste(year,month1,day,time,sep=""))
childline1 <- childline %>%
mutate(date1=as.POSIXlt(childline$date))
#creating weekdays and months variables
childline2 <- childline1%>%
mutate(month=months(childline1$date1))%>%
mutate(wkday=weekdays(childline1$date1))
#creating numerical weekdays variables
children <- childline2[,c(1,seq(20,29,by=1))]
children$wkdaynum <- case_when(
children$wkday=="Monday" ~ 1,
children$wkday=="Tuesday" ~ 2,
children$wkday=="Wednesday" ~ 3,
children$wkday=="Thursday" ~ 4,
children$wkday=="Friday" ~ 5,
children$wkday=="Saturday" ~ 6,
children$wkday=="Sunday" ~ 7
)

#counts as per day, time, weekday and months
children1 <- children %>%count(date1, wkday,month, month1)
children2 <- children %>%count(time_index, date1, wkday,month, month1, dttime)
children3 <- children %>%count(time)
children4 <- children %>%count(wkday,wkdaynum)
children5 <- children %>%count(month, month1)
children6 <- children %>%count(day)
```

```
#date <- seq(as.POSIXlt("2016-01-01"),as.POSIXlt("2016-07-12"), by="day")
#creating a date time sequence similar to our data
dttime <- seq(as.POSIXlt("2016-01-01 07:00:00"),as.POSIXlt("2016-07-12 23:00:00
"), by="hour")
wkday <- weekdays(dttime)
month <- months(dttime)
month1 <- trimws(substr(dttime,6,7))
#merging the sequence with our data
fulldata <- data.frame(dttime,wkday,month, month1)
merged <- left_join(fulldata,children2,by=c("dttime"="dttime", "wkday"="wkday", "mo
#using package zoo for missing intervals
zoovalues <- zoo(merged$n,merged$dttime)
set.seed(11111111)
approx <- as.vector(round(na.approx(zoovalues)))
fulldata2 <-data.frame(dttime,wkday,month, approx,merged$n, month1)
fulldata2$wkdaynum <- case_when(
fulldata2$wkday=="Monday" ~ 1,
fulldata2$wkday=="Tuesday" ~ 2,
fulldata2$wkday=="Wednesday" ~ 3,
fulldata2$wkday=="Thursday" ~ 4,
fulldata2$wkday=="Friday" ~ 5,
fulldata2$wkday=="Saturday" ~ 6,
fulldata2$wkday=="Sunday" ~ 7
)
dataapprox <- fulldata2[,c(1,4)]
#creating time series data using zoo or ts to plot run sequential
childrents <- zoo(dataapprox$approx, dataapprox$dttime)
childrents1 <- ts(approx, freq=24)
pdf(file = "outputs/Run Sequential.pdf")
plot(childrents, pch=18, col="black",main = "Childline Call Frequency Time series p
```

```
dev.off()
#acf of the timeseries data
pdf(file = "outputs/acf1.pdf")
acf(childrents1)
dev.off()
#PACF of the timeseries data
pdf(file = "outputs/pacf1.pdf")
pacf(childrents1)
dev.off()
# Hourly line plot
pdf(file = "outputs/line.pdf")
ggplot(data=children3, aes(x=time, y=n, group=1)) +
geom_line(linetype = "dashed")+

geom_point()+
ggtitle("Call Frequency Grouped per Hour") +
xlab("Hour") + ylab("Call Frequency")
dev.off()
#weekly lineplot
pdf(file = "outputs/line2.pdf")
children4_ <- children4[order(children4$wkdaynum),]
ggplot(data=children4_, aes(x=wkdaynum, y=n, group=1)) +
geom_line(linetype = "dashed")+
geom_point()+
ggtitle("Call Frequency Grouped per Weekday") +
xlab("Weekday") + ylab("Call Frequency")
dev.off()
#lagplot
x1<-children2$n[1:length(children2$n)-1]
x2<-children2$n[2:length(children2$n)]
```

```
data1<-data.frame(x1,x2)
pdf(file = "outputs/lag.pdf")
ggplot(data1,aes(x=x2,y=x1)) + geom_point()+
geom_smooth(method=lm, se=FALSE) + ggtitle("Lag Plot of Intraday Call frequency vs
xlab("Lag 1") + ylab("Call Frequency")
#cor(data1)
dev.off()
#residual analysis
arfit=lm(x1~x2)#Does regression, stores results object named ar1fit
summary(arfit) # This lists the regression results
pdf(file = "outputs/acfres2.pdf")
acf(arfit$residuals, xlim=c(1,18)) # ACF of the residuals for lags 1 to 18
dev.off()
#residual analysis
pdf(file = "outputs/resplot.pdf")
plot(arfit$fit,arfit$residuals, xlab="Fitted Value", ylab="Residuals", main="Plot c
abline(h=0)
dev.off()
#decomposition as hourly data
pdf(file = "outputs/decomp1.pdf")
childrents1 %>%
tail(24*7*4) %>%
decompose() %>%
autoplot()
dev.off()
#decomposition as with multiple periods
childrenmsts<-approx %>% msts( seasonal.periods = c(24, 24*7))
pdf(file = "outputs/decomp2.pdf")
childrenmsts %>% head( 24 *7 *4) %>% mstl() %>% autoplot()
dev.off()
```

```
#plot(childrenmsts)
#log(chidlrenmsts)
#log transformed
logchildren <- log(childrents1)
pdf(file = "outputs/Run Sequential log.pdf")
plot(logchildren,type="o", pch=18, col="green",main = "Log transformed Childline Ca
dev.off()

#weekly and daily differenced
#ldchildren <- diff(logchildren, differences=1)
ts.diff.week = diff(childrents1, 24*7,1)
ts.diff.day = diff(ts.diff.week, 24)
pdf(file = "outputs/Run Sequential diff.pdf")
plot(ts.diff.day,type="o", pch=18, col="blue",main="Differenced Call Frequency",yla
dev.off()
#acf and pacf plot
par(mfrow=c(2,2))
pdf(file = "outputs/acfpacfarima.pdf")
Acf <- forecast::Acf
ggPacf(ts(ts.diff.week, frequency = 24), lag.max=24*7*28, col="#444444")+scale_x_co
ggAcf(ts(ts.diff.week, frequency = 24), lag.max=24*7*28, col="#444444")+scale_x_con
ggPacf(ts(ts.diff.day, frequency = 24), lag.max=24*6, col="#444444")+scale_x_contin
ggAcf(ts(ts.diff.day, frequency = 24), lag.max=24*6, col="#444444")+scale_x_continu
dev.off()
#stationarity test
adf.test(ts.diff.day)
kpss.test(ts.diff.day, null="Trend")
set.seed (1000001)
#Crossvalidation
len<-28 #no of weeks between jan - 12july
```

```
length1=length(childrents1)
len1<- round(0.8*len*24*7, digits=0)
train <- childrents1[1:len1]
test <- childrents1[(len1+1):length1]
#fit various potential arima models
library(smooth)
model.msar1=msarima(train, orders=list(ar=c(1,0,5),i=c(0,0,0),ma=c(1,1,1)),lags=c(1
model.msar2=msarima(train, orders=list(ar=c(1,0,5),i=c(0,0,0),ma=c(1,1,0)),lags=c(1
model.msar3=msarima(train, orders=list(ar=c(1,0,4),i=c(0,0,0),ma=c(1,1,1)),lags=c(1
model.msar4=msarima(train, orders=list(ar=c(1,0,4),i=c(0,0,0),ma=c(1,1,0)),lags=c(1
model.ar1=auto.arima(train)
arics=cbind(round(model.ar1[["aic"]],digits=0),
round(model.ar1[["aicc"]],digits=0),
round(model.ar1[["bic"]],digits=0), "-")
ics <- as.data.frame(rbind(round(model.msar1[["ICs"]],digits=0),
round(model.msar2[["ICs"]],digits=0),
round(model.msar3[["ICs"]],digits=0),
round(model.msar4[["ICs"]],digits=0),
arics))
row.names(ics)=c(model.msar1[["model"]],
model.msar2[["model"]],
model.msar3[["model"]],
model.msar4[["model"]],
"ARIMA(4,1,2)")
kable(ics, caption = "IC(s) of each possible ARIMA model")%>%
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T) %>%
row_spec(1, bold = T, color = "white", background = "#D7261E")

ics1=knitr::kable(ics, caption = "IC(s) of each possible ARIMA model","latex")%>%
```

```
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T) %>%
row_spec(1, bold = T, color = "white", background = "#D7261E")
writeLines(ics1, 'accs.tex', file.path("C:/Users/grace/OneDrive/Documents/2. RESEAR
#residual analysis
pdf(file = "outputs/checkresiduals.pdf")
checkresiduals(model.msar1)
dev.off()
#residual analysis
Box.test(model.msar1$residuals, type="Ljung-Box")
pdf(file = "outputs/acfres.pdf")
acf(model.msar1$residuals)
dev.off()
set.seed(111111)
#36 days ahead:
#model1
for_msar1.30 = forecast(model.msar1, h=886)
y_hat_2_msar1.30=for_msar1.30$mean
pdf(file = "outputs/formsar30.pdf")
plot(for_msar1.30)
dev.off()
msar_1.30=accuracy(for_msar1.30$mean,test)
#model4
for_msar4.30 = forecast(model.msar4, h=886)
y_hat_2_msar4.30=for_msar4.30$mean
plot(for_msar4.30)
msar_4.30=accuracy(for_msar4.30$mean,test)
set.seed(111111)
#7 days ahead:
#model1
```

```
for_msar1.7 = forecast(model.msar1, h=168)
y_hat_2_msar1.7=for_msar1.7$mean
pdf(file = "outputs/formsar7.pdf")
plot(for_msar1.7)
msar_1.7=accuracy(for_msar1.7$mean,test[1:168])
dev.off()
#model4
for_msar4.7 = forecast(model.msar4, h=168)
y_hat_2_msar4.7=for_msar4.7$mean
plot(for_msar4.7)
msar_4.7=accuracy(for_msar4.7$mean,test[1:168])
acc_msa_30= as.data.frame(rbind(msar_1.30, msar_4.30))
row.names(acc_msa_30)=c(paste("30 days: ", model.msar1[["model"]]), paste("30 days:
acc_msa_7= as.data.frame(rbind(msar_1.7, msar_4.7))
row.names(acc_msa_7)=c(paste("7 days: ", model.msar1[["model"]]), paste("7 days: ",
msar=acc_msa_7

row.names(msar)=NULL
msar$day=7
msar$model="msar"
for (i in 1:36) {
j=i*24
test_i=test[1:j]
msar_i = forecast(model.msar1, h=j)
accrms =c(as.numeric(accuracy(msar_i$mean,test_i)),i,"SARIMA")
msar[nrow(msar)+1,]=accrms
}
#Fit NNAR model
set.seed(111111)
train1 <- childrenmsts[1:len1] %>% msts( seasonal.periods = c(24, 24*7))
```

```
model.ann=nnetar(train1)
model.ann
set.seed(111111)
#36 days ahead:
library(tibble)
library(sweep)
detach("package:smooth", unload=TRUE)
detach("package:forecast", unload=TRUE)
library(forecast)
for_ann30 = forecast(model.ann, h=886)
y_hat2_ann30=for_ann30$mean
acc_ann_30=as.data.frame(accuracy(for_ann30$mean,test))
row.names(acc_ann_30)="30 days: NNAR (35,1,18)[168]"
set.seed(111111)
for_ann7 = forecast(model.ann, h=168)
y_hat2_ann7=for_ann7$mean
acc_ann_7=as.data.frame(accuracy(for_ann7$mean,test[1:168]))
row.names(acc_ann_7)="7 days: NNAR(35,1,18)[168]"
pdf(file = "outputs/forann7.pdf")
plot(for_ann7)
dev.off()
pdf(file = "outputs/forann30.pdf")
plot(for_ann30)
dev.off()
#Hybrid model
#training the residuals
res_msar1=model.msar1$residuals
res=length(res_msar1)
len2<- round(0.8*res, digits=0)
anntrain1 <- res_msar1[1:len2] %>% msts( seasonal.periods = c(24, 24*7))
```

```
anntest1 <- res_msar1[(len2+1):res]
res_msar4=model.msar4$residuals
annlength=length(anntest1)
hy11=nnetar(anntrain1, P=5, p=1,size=1, repeats=40)
for_11= forecast(hy11, h=annlength)
accsh11=accuracy(for_11$mean,anntest1)

hy12=nnetar(anntrain1, P=5, p=3,size=1, repeats=40)
for_12= forecast(hy12, h=annlength)
accsh12=accuracy(for_12$mean,anntest1)
hy13=nnetar(anntrain1, P=6, p=2,size=1, repeats=40)
for_13= forecast(hy13, h=annlength)
accsh13=accuracy(for_13$mean,anntest1)
hy14=nnetar(anntrain1, P=4, p=2,size=1, repeats=40)
for_14= forecast(hy14, h=annlength)
accsh14=accuracy(for_14$mean,anntest1)
hy15=nnetar(anntrain1, P=5, p=2,size=3, repeats=40)
for_15= forecast(hy15, h=annlength)
accsh15=accuracy(for_15$mean,anntest1)
hy16=nnetar(anntrain1, P=5, p=2,size=1, repeats=30)
for_16= forecast(hy16, h=annlength)
accsh16=accuracy(for_16$mean,anntest1)
hy17=nnetar(anntrain1, P=5, p=2,size=1, repeats=20)
for_17= forecast(hy17, h=annlength)
accsh17=accuracy(for_17$mean,anntest1)
hy18=nnetar(anntrain1, P=5, p=2,size=1, repeats=40)
for_18= forecast(hy18, h=annlength)
accsh18=accuracy(for_18$mean,anntest1)
hy19=nnetar(anntrain1, P=5, p=2,size=2, repeats=40)
for_19= forecast(hy19, h=annlength)
```

```
accsh19=accuracy(for_19$mean,anntest1)
hy19l=nnetar(anntrain1, P=5, p=2,size=1,decay=0.1, repeats=30)
for_19l= forecast(hy19l, h=annlength)
accsh19l=accuracy(for_19l$mean,anntest1)
hybtesting=rbind(accsh17, accsh18, accsh16, accsh11, accsh12, accsh13, accsh14, acc
row.names(hybtesting)=c(paste(hy17$method,"repeats=20"),
paste(hy18$method,"repeats=40"),
paste(hy16$method,"repeats=30"),
paste(hy11$method,"repeats=30"),
paste(hy12$method,"repeats=40"),
paste(hy13$method,"repeats=40"),
paste(hy14$method,"repeats=40"), paste(hy19$method,"repeats=40"),
paste(hy15$method,"repeats=40"))
kable(hybtesting, digits=2, caption = "Training the Hybrid-ANN parameters")%>%
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T) %>%
row_spec(2, bold = T, color = "white", background = "#D7261E")
hybst=kable(hybtesting, digits=2, caption = "Training the Hybrid-ANN parameters", "
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T) %>%
row_spec(2, bold = T, color = "white", background = "#D7261E")
#ARIMA MODEL FORECAST+ ANN MODEL FORECAST with the two best ARIMA and ANN models
#hybrid model
set.seed(111111)
#model1
model.annhy1=nnetar((res_msar1 %>% msts( seasonal.periods = c(24, 24*7))), P=5,size
for_annhy1.30 = forecast(model.annhy1, h=886)
y_hat_2_annhy1=for_annhy1.30 $mean

h1=as.numeric( y_hat_2_annhy1)
```

```
h2=as.numeric(y_hat_2_msar1.30)
y_hat_2_comb1.30=h1+h2
acchy1.30=accuracy(y_hat_2_comb1.30,test)
#model4
model.annhy4=nnetar((res_msar4 %>% msts( seasonal.periods = c(24, 24*7))),p=2,size=
for_annhy4.30 = forecast(model.annhy4, h=886)
yarima=as.numeric(y_hat_2_msar4.30)
y_hat_2_annhy4=for_annhy4.30$mean
h3=as.numeric( y_hat_2_annhy4)
h4=yarima
y_hat_2_comb4.30=h3+h4
acchy4.30=accuracy(y_hat_2_comb4.30,test)
acc_hy_30=as.data.frame(rbind(acchy1.30,acchy4.30))
row.names(acc_hy_30)=c(paste("30 days: ", model.msar1[["model"]],model.annhy1$metho
acc_30=rbind(acc_msa_30,acc_ann_30,acc_hy_30)
best_30= as.data.frame(rbind(msar_1.30,acc_ann_30,acchy1.30))
row.names(best_30)=c(paste("30 days: ", model.msar1[["model"]]),
paste("30 days: ", model.ann$method),
paste("30 days: Hybrid", model.msar4[["model"]],"-", model.annhy1$method))
kable(best_30, digits=2, caption = "36 days forecast: Metrics of the Models")%>%
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T) %>%
row_spec(3, bold = T, color = "white", background = "#D7261E")
best_30.1=best_30[,c(2,3,5)]
b30=best_30.1%>%dplyr::select(MAE) %>%
kable(digits=2, caption = "36 days forecast:Accuracy Metrics of the Models", "latex
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T)
#model1
for_annhy1.7 = forecast(model.annhy1, h=168)
```

```
y_hat_2_annhy1.7=for_annhy1.7$mean
h5=as.numeric( y_hat_2_annhy1.7)
h6=as.numeric(y_hat_2_msar1.7)
y_hat_2_comb1.7=h5+h6
acchy1.7=accuracy(y_hat_2_comb1.7,test[1:168])
#model4
for_annhy4.7 = forecast(model.annhy4, h=168)
y_hat_2_annhy4.7=for_annhy4.7$mean
h7=as.numeric( y_hat_2_annhy4.7)
h8=as.numeric(y_hat_2_msar4.7)
y_hat_2_comb4.7=h7+h8
acchy4.7=accuracy(y_hat_2_comb4.7,test[1:168])

acc_hy_7=as.data.frame(rbind(acchy1.7,acchy4.7))
row.names(acc_hy_7)=c(paste("7 days: ", model.msar1[["model"]],model.ann$method), p
acc_7=rbind(acc_msa_7,acc_ann_7,acc_hy_7)
best_7= as.data.frame(rbind(msar_1.7,acc_ann_7,acchy1.7))
row.names(best_7)=c(paste("7 days: ", model.msar1[["model"]]),
paste("7 days: ",model.ann$method),
paste("7 days: Hybrid", model.msar4[["model"]], "-", model.annhy1$method))
kable(best_7, digits=2, caption = "7 days forecast:Accuracy Metrics of the Models")
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T)
#best_7.1=best_7[,c(2,5)]
best_7.1=best_7[,c(2,3,5)]
b7=best_7.1%>%dplyr::select(MAE) %>%
kable(digits=2, caption = "7 days forecast:Accuracy Metrics of the Models", "latex"
kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
column_spec(1, bold = T)
TEST=as.ts(test,frequency=24)
```

```
pdf(file = "outputs/forhy30.pdf")
autoplot(TEST) +
autolayer(as.ts(y_hat_2_comb1.30,frequency=24), series = 'HYBRID')
dev.off()
TEST7=as.ts(test[1:168],frequency=24)
pdf(file = "outputs/forhy7.pdf")
autoplot(TEST7) +
autolayer(as.ts(y_hat_2_comb1.7,frequency=24), series = 'HYBRID')
dev.off()
pdf(file = "outputs/forcm30.pdf")
autoplot(TEST) +
autolayer(as.ts(y_hat_2_comb1.30,frequency=24), series = 'HYBRID') +
autolayer( as.ts(as.numeric(y_hat_2_msar1.30),frequency=24), series = 'SARIMA') +
autolayer( as.ts(as.numeric(y_hat2_ann30),frequency=24), series = 'NNAR')
dev.off()
pdf(file = "outputs/forcm7.pdf")
autoplot(TEST7) +
autolayer(as.ts(y_hat_2_comb1.7,frequency=24), series = 'HYBRID') +
autolayer( as.ts(as.numeric(y_hat_2_msar1.7),frequency=24), series = 'SARIMA') +
autolayer( as.ts(as.numeric(y_hat2_ann7),frequency=24), series = 'NNAR')
dev.off()
set.seed(1111111)
anns=acc_ann_7
row.names(anns)=NULL
anns$day=7
anns$model="ann"
for (i in 1:36) {
j=24*i

test_i=test[1:j]
```

```
mann_i = forecast(model.ann, h=j)
accr_ann =c(accuracy(mann_i$mean,test_i),i,"NNAR")
anns[nrow(anns)+1,]=accr_ann
}
set.seed(111111)
hys=acc_ann_7
row.names(hys)=NULL
hys$day=7
hys$model="ann"
yarima=as.numeric(y_hat_2_msar1.30)
for (i in 1:36) {
j=i*24
test_i=test[1:j]
mannhy_i = forecast(model.annhy1, h=j)
yi=mannhy_i$mean
hi_ann=as.numeric(yi)
hi_msa=yarima[1:j]
y_combi=hi_ann+hi_msa
accrhy =c(as.numeric(accuracy(y_combi,test_i)),i,"HYBRID")
hys[nrow(hys)+1,]=accrhy
}
results=rbind(hys[2:37,],anns[2:37,], msar[3:38,])
results$day1=round(as.numeric(results$day),digits=2)
results$MAPE1=round(as.numeric(results$MAPE),digits=2)
results$MAE1=round(as.numeric(results$MAE),digits=2)
results$RMSE1=round(as.numeric(results$RMSE),digits=2)
pdf(file = "outputs/linemape.pdf")
ggplot( data=results,aes(x=day1, y=MAPE1, group=model, color=model)) +
geom_line()
dev.off()
```

```
pdf(file = "outputs/linermse.pdf")
ggplot( data=results, aes(x=day1, y=RMSE1, group=model, color=model)) +
geom_line()
dev.off()
pdf(file = "outputs/linemae.pdf")
ggplot( data=results, aes(x=day1, y=MAE1, group=model, color=model)) +
geom_line()
dev.off()
residuals_msar=test-as.numeric(y_hat_2_msar1.30)
residuals_ann=test-as.numeric(y_hat2_ann30)
residuals_hy=test-y_hat_2_comb1.30
dm.test(residuals_msar,residuals_hy,alternative = "less",h=886,power=2 )
dm.test(residuals_ann,residuals_hy,alternative = "less",h=886,power=2 )
```

Appendix B

Turnitin Report

Curiginal

Document Information

Analyzed document	136947 GRACE WAIRIMU - THESIS V3.pdf (D138131024)
Submitted	2022-05-26T18:25:00.0000000
Submitted by	
Submitter email	Grace.WangOmbe@strathmore.edu
Similarity	2%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

W	URL: https://www.science.gov/topicpages/a/arima-based+time+series.html Fetched: 2022-05-26T18:25:33.8600000	1
SA	**BScThesis_MMvanderMeulen_2639613.pdf** Document BScThesis_MMvanderMeulen_2639613.pdf (D110192236)	5
SA	**CAnners.pdf** Document CAnners.pdf (D30141368)	2
SA	**Master_s Thesis Fredrik Nikolaisen Sävås.pdf** Document Master_s Thesis Fredrik Nikolaisen Sävås.pdf (D8410549)	2
W	URL: https://www.researchgate.net/publication/318418989_Hybrid_ARIMA-ANN_Modelling_for_Forecasting_the_Price_of_Robusta_Coffee_in_India Fetched: 2021-03-02T04:25:29.0870000	1
SA	**12626509.pdf** Document 12626509.pdf (D50365365)	1
W	URL: https://bioinfopublication.org/include/download.php?id=BIA0003536 Fetched: 2022-05-26T18:25:33.9630000	1
SA	**732A64 Master's Thesis mowas455.pdf** Document 732A64 Master's Thesis mowas455.pdf (D138121236)	1
W	URL: https://pdfs.semanticscholar.org/eda0/4db30cfd1b7d0e525f80c6775b3167ecf54e.pdf Fetched: 2021-11-06T20:46:07.1830000	1
W	URL: https://www.mathsjournal.com/pdf/2020/vol5issue2/PartA/5-1-13-832.pdf Fetched: 2021-12-27T12:23:40.1200000	1

Curiginal

Entire Document

A Comparative study of Hybrid Neural Network and ARIMA Models with Application to Forecasting Intra-day Child-line Calls in Kenya. Grace Wairimu Wang'ombe Submitted in total fulfilment of the requirements for the degree of Masters of Science in Statistical Science of Strathmore University Institute of Mathematical Sciences Strathmore University Nairobi, Kenya

March 2022 This thesis is available for Library use through open access on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement. ii

Declaration I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself. © No part of this thesis may be reproduced without the permission of the author and Strathmore University. Name: .... Grace Wairimu Wang'ombe .... Signature: .... Date: .... May 26, 2022 .... Approval The thesis of Grace Wairimu Wang'ombe was reviewed and approved by the following: Dr. Collins Odhiambo Supervisor, Institute of Mathematical Sciences, Strathmore University. Dr. Thaddeus Wandera Egondi Supervisor, Institute of Mathematical Sciences, Strathmore University. iii

Appendix C

Ethical Approval

Strathmore UNIVERSITY

27th June 2022

Ms Wang'ombe Grace,
grace.wangombe@strathmore.edu

Dear Ms Wang'ombe,

RE: Forecasting Intraday Childline Calls in Kenya using a Hybrid of Neural Network and ARIMA models

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU Masters'** research proposal. Your application reference number is **SU-IERC1368/22.** The approval period is **27th June 2022 to 26th June 2023.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-ISERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Prof Fred Were,
Chairperson; SU-ISERC