AIMS | African Institute for Mathematical Sciences
CAMEROON

Statistical Methods for Predicting Prison Populations

Blessing Itoro Bassey (blessing.bassey@aims-cameroon.org)
African Institute for Mathematical Sciences (AIMS)
Cameroon

Supervised by: Sam Cuthbertson
UK Civil Service

31 May 2019

Submitted in Partial Fulfillment of a Structured Masters Degree at AIMS-Cameroon

Abstract

This study aims to use statistical methods to predict the prison population. Using data collected on-line at [14] from England and Wales, the prison population data were categorized into different sections but for the purpose of this work, we considered the prison population whose sentence length is 4 years or more (excluding the indeterminate sentence). To develop a model for prediction, three models were considered. First, the regression model; second, the SARIMA model (0,2,1)(0,0,1)[12]; and lastly, the stock-flow model.

The regression model did not produce a good fit as a result of the seasonal and trend patterns in time, hence, the introduction of the natural cubic spline accounted for it. Diagnostic checking was carried out on the regression model with the natural cubic spline, the assumption of the population error term were met, and the QQ-plot also followed a normal distribution. For the time series analysis, an additive decomposition was made which also revealed the presence of seasonality and trend, hence, an auto.arima with frequency 12 was used to generate a SARIMA model (0,2,1)(0,0,1)[12]. The model was further validated using the Ljung-Box test with a p-value of 0.8562, which is highly significant, meaning that the model does not show a lack of fit. Also, the ACF and PACF of the residual plot had all the lags within the 95% confidence interval. Finally, using the stock flow model to estimate the inflow of offenders, their releases as well as the closing stock for a particular period of time.

For an accuracy measure, we compared the test set (prison population from 2011-2016) with the validation set (prison population of 2017), and afterward, the forecast for the prison population for January 2018 to March 2019 was made using the three models, see Table 4.14. The SARIMA model had the smallest RMSE, of 231, the regression model which had 396, and the stock flow had 5795. The simple stock-flow was further developed by introducing recidivism. Incorporating other factors like recidivism will be useful for better predictions. Hence, it is practicable to recommend the SARIMA (0,2,1)(0,0,1)[12] model for predicting the prisons population of sentence length of 4 years or more.

Keywords: Time Series, Forecasting, Regression, Stock Flow, Prison Population, Recidivism.

Declaration

I, the undersigned, hereby declare that the work contained in this essay is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Blessing Bassey, 31 May 2019.

Contents

1. Introduction

Prison, which is also known as a place for correction or detention center is a facility in which inmates are forcibly kept and denied a variety of freedoms under authority with the purpose highlighted below [16].

- *Retribution :-* This is the penalty or sanction on criminals for unlawful acts against society
- *Incapacitation :-* Removal of criminals from society so that they can no longer harm innocent people
- *Recuperation :-* Designing activities to reform or rehabilitate criminals into law-abiding citizens.

This project will analyze extensive published data on the prison population to identify the need for future prison places having known the estimated prisons populations, and also aims to demonstrate the role statistics can play in a real-world policy problem, and the importance of making recommendations based on evidence.

1.1 Problem and Summary of data collected from England and Wales

In 2017, there were 86,000 people in prison in England and Wales. Whilst this is an increase of 40% compared to twenty years ago, these figures have barely changed in the last five years. This project will analyse extensive published data on the prison population to identify the need for future prison places. This data was collected online at [14] using data from England and Wales as a start up point, the skills would be highly transferable to solving other problems. The prison population data is categorized into different sections; (1) The prison population by type of custody age group and sex, i.e determinate or indeterminate sentence and also the sentenced length(if it's determinate). (2) The prison population by offense group i.e Fraud, robbery, sexual offense etc. (3) Prison population by ethnic group and sex as well as their admission, sentence length, and release. (4) The prison population by religion and national status as well as their individual admission sentence length and release date. The data were collected quarterly, i.e on January 31st 2018, April 30th 2018 e.t.c.

1.2 Characteristics of prison populations

1.2.1 Sentencing. This refers to the punishment ordered by a trial court leading to instant imprisonment. This can be divided into 2 categories.

- Determinate sentences:- This is the most common type of prison sentence whereby the court sets a fixed length or the maximum period of time the offender spends in the prison. e.g Standard determinate sentence, Extended Determinate Sentence etc.
- Indeterminate sentences:- This is the sentence imposed for a crime without a given definite duration, i.e individual on whom such sentences are imposed has no guaranteed release date or time. Examples are, **Whole life sentence** that is, an imprisonment without the possibility of parole, which means that release might not be considered; Imprisonment for Public Protection (IPP) sentence, which is mainly for offenders of age 18 and above who has committed a serious crime e.g violent or sexual offense, for which the maximum penalty is 10 years and above. This sentence

was introduced by Section 225 of the Criminal Justice Act 2003 which took effect from 2005 and then was abolished in 2012 [13].

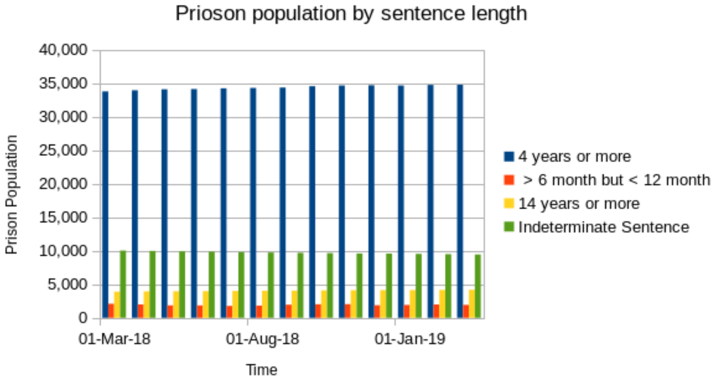


Figure 1.1: Prison Population by Sentence length

Figure 1.1 displays the male and female prison population under immediate custodial by sentence length. From the graph it is observed that the sentence length of "4 years or more" had the highest prison population, while the sentence length of "6 months but less than 12 months" had the least prison population.

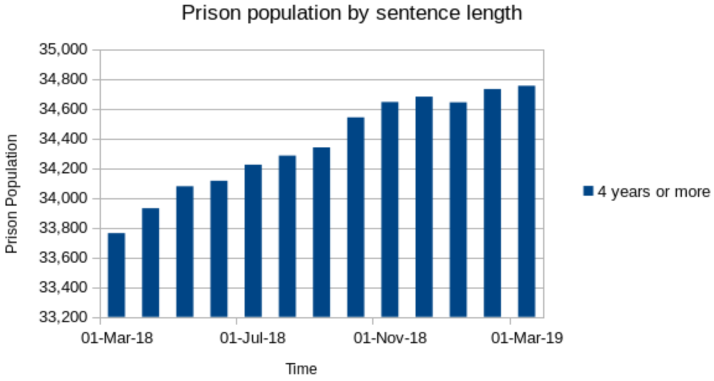


Figure 1.2: Prison population of Sentence length of 4 years or more

The prison population of sentence length of "4 years or more" is displayed in Figure 1.2. It is observed that the population increased with time. As at May 2018, the population was approximately 33,750 but by May 2019 the population had increased to about 34,750.

1.3 The Time Series Analysis

When faced with real-life instances and we ask questions like what will be the stock price after a month? What will be the air temperature tomorrow? What will the population of this country be by next year?

One common factor in all these questions is **time**. Therefore, data recorded on a timely basis is called time series data, and the different methods for analyzing this type of data in order to extract meaningful statistics understands the past as well as predicts the future. In time series analysis, there are different types of model to choose from. This will be discussed fully in the next chapter.

1.4 The Stock Flow model

There is a need to go beyond just simply analyzing and visualizing the formation of the prison population, a more powerful procedure or approach that enables us to create a better representative of the system will allow us to explore its behavior and to test the effect of changes to the system's formation and the policies governing its behavior.

(a) **Stocks**:- This represents a part of a system whose worth at any given instant of time is a function of the system's past behavior. This value cannot just be determined by measuring the value of other parts of the system at that particular time, one way to solve this is by calculating how it changes at every point in time and adding up these changes [22].

(b) **Flows**:- This represents the rate at which stock changes at a particular time, an in-flow into the stock causes the stock to increase while an outflow from the stock causes the stock to decrease or flow out of a stock (causing it to decrease).

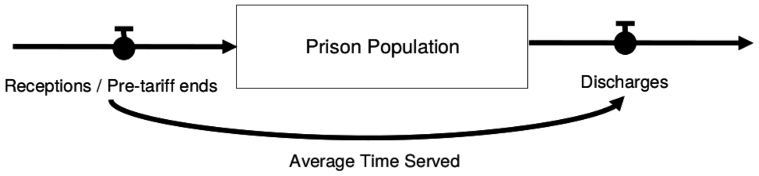


Figure 1.3: A stock flow diagram of prisons population [17]

From Figure 1.3, if the reception is greater than the discharge, then the prison population increases leading to overcrowding, on the other hand, if the discharge is greater than the reception, the prison population is reduced. Understanding the variety of the prison population is important to what may be needed in the future. There are three levers that society can pull (through policy decisions by the Department of Justice and the courts) [21]. How many offenders are sent to prison (Admission), How long they stay (Sentence length), and How many were set free (Released).

1.5 Roadmap Of Study

Chapter 1 involves the general introduction which includes the objectives of the study and layout of the essay. In chapter 2, we would discuss the literature review of previous study on predictions. Chapter 3 involves the statistical methods for predicting prison population; the Regression model, Time Series model, and the Stock flow model. In chapter 4, the analysis and Results would be discussed and finally in chapter 5 we give the conclusion and recommendations.

2. Literature Review

2.1 The Need for Making Predictions

Constantly, we consciously or unconsciously predict the future or forecast what might happen, be it a short term or long term prediction. With this, we have a better chance to control things thereby helping us make good decisions in accomplishing our set goals and objectives. The ability to predict is finding the connection between cause and effect. If we can connect the cause of today to the effect that might happen tomorrow, then we can predict, with this chain, we can give the forecast of what will happen next week, next month and even years to come. [18]. In this chapter, we will review the literature of some of the studies done on prediction or forecast, particularly, studies with a focus on population prediction.

Sandra Evans Skovron [12] assessed public attitudes towards reducing prison crowding through telephone surveys of adult residents of two major cities: Cincinnati and Columbus, Ohio. The support from the public was substantial as it was a community-based correction and incentive. Prison construction received only moderate support while high levels of public disapproval were found for shortening sentences and increasing parole board authority which was the major cause of prison crowding. The Probit regression analysis that was conducted on the relationship between support for policies in other to reduce prison crowding and respondent characteristics revealed that attitudinal variables were more consistently related to public opinion than were demographic variables, that is, there was a need for more prison place in other to reduce prison crowding.

Sarah Armstrong from the University of Glasgow alongside Elizabeth Fraser from the Scottish Government analytical services [19] projected the prison population. They projected a range of time periods in order to account for changes in trend over time using the time series analysis based on linear regression and exponential smoothing. Six variants reflecting the overall trends over the 10 years, 25 years and 40 years term were considered, which best reflects the current situation and the need to compensate for inherent passing off of the smaller groups with time.

The challenges facing prison managers include questions regarding excessive reliance on imprisonment as punishment, the appropriateness of the inmates for whom imprisonment is imposed, problems of crowding, the effectiveness of imprisonment, and the cost of imprisonments. As a result of over-capacitated coupled with poor ventilation in prison places, tuberculosis (TB) continues to spread in South African Prison [10]. The Department of Mathematics and Applied Mathematics, University of the Western Cape, proposed a system dynamic model of demand and supply to describe the population dynamics of Tuberculosis disease in prisons. This considers the inflow of susceptible into the prison population and a non-inflow of infected persons into the prison, which will ensure that Tuberculosis can be eradicated from the prison population. By the simulations generated with the model, it illustrates how it can be useful in making future projections of the levels of prevalence of Tuberculosis disease, and to quantify the effect of interventions such as screening, treatment or reduction of transmission parameter values through improved living conditions for inmates. This model is particularly useful to the World Health Organization and the governments, for reduction of Tuberculosis disease prevalence and ultimately eradicating it.

2.2 Application of Stock Flow Model

Many models can be modified with the mind of tracking changes in the location or abundance (the flow) of one or more "things" (the stock). i.e The system dynamic modeling. The concept of Stock and flow model is applied in many fields such as the financial system, flows of produced goods and services through the real economy, and flows of physical materials through the natural environment. The study of the CJ Special Interest Group, [15] applied the concept of Stock flow in Forecasting and Model Development, stating that flows are important because they carry cost implications for the National Offender Management Service and because they are needed for forecasting the admission of an inmate, the sentence length, and the release.

The Stock-Flow Model [9] for Forecasting Labor Supply was developed by Chun-pong Sing, and he did this to estimate the supply of labor in the construction industry. The findings were determined using a stock-flow model, which enabled the determination of future aging distribution trends and workforce supply for specific trade types. The inflow of labor is gotten from the idea of transition such as new entrants, promotion, or transfers etc while the promotion to a specific organization or position generates the outflow. The developed stock-flow model can be applied in countries where registration schemes for construction workers are in use.

Fluctuation in housing prices is an extensive circumstance or occurrence which has been repeatedly observed in many countries. [11] Elizabeth Steiner used a Stock-Flow Model for Housing Market in Switzerland. With the model, she was able to analyze the major determinants of the Swiss housing by quantifying the gap existing between the previous supply and the desired level of housing stock, also to measure the effect of these housing market fluctuations on prices. The model was constructed in such a way that it was possible to simulate different scenarios in order to forecast potential developments on the housing market. The demand and supply for housing stock determine the Housing prices and also influences the flow of residential investment. Higher housing prices gives rise to the need of building new houses and boost residential investment.

England and Wales have one of the highest incarceration rates in Western Europe. As at December 2007, the prison population was 1.53 for every 100,000 members of the population in England and Wales. previous studies have shown that population can be predicted using various models, but for the purpose of this study, We will be predicting the prison population of England and Wales (sentence length of 4 years or more) using one of the three models that would be considered in subsequent chapters.

3. Methodology

3.1 Research Design

In this study, three methods would be used to predict the prison population, Regression model, Time Series model and the Stock Flow model.

3.2 The Time Series Model

A time plot or a time series graph basically displays values against time. Although it is similar to the x-y graphs, while an x-y graph plots varieties of "x" variables e.g price, age, heights etc, a time plots only displays time on the x-axis. The aim of time series analysis is to identify patterns in the data and then use the data for predictions. Before we move on, we need to discuss some important components described by most time series data and these include trend, seasonality, and noise [23].

1. **Trend:**- This is a variation that moves up or down in a reasonably predictable pattern over time and does not repeat
2. **Seasonality:**- A systematic linear or nonlinear characteristic of a time series in which the data experiences regular and predictable pattern that recurs or repeats over a one-year period.
3. **Noise:**- A time series data is white noise if the variables are independent and identically distributed with a mean of zero. This implies that each value has a zero correlation with all other values in the series, and that all variables have the same variance.

3.3 Stationarity in Time Series

Stationarity is when the statistical properties of a process generating a time series do not change over time, although it does not mean that the series does not change over time, it is just that the way it changes does not itself change over time, for example, let's consider a linear function, $y = 2x$, the value of y changes as x grows, but the way it changes remains constant. Hence, in time series a stationary process has the property of the mean, variance and autocorrelation structure being constant, i.e they do not change over time. Not all time series data are stationary, but for proper prediction and forecast, we need a stationary time series data.

Assumptions:-

1. Constant mean: $\mathbb{E}(y_t) = \mu, \qquad \forall t.$
2. Constant variance: $var(y_t) = \gamma(0) = \sigma^2, \qquad \forall t$
3. Non correlated: $Cor(y_{t_1}, y_{t_{t-1}}) = \gamma[t_1, t_{t-1}], \qquad \forall t$

3.4 To Check for Nonstationarity

1. **Run sequence Plots:-** Plotting a run sequence plot to see if there is anything like trends or seasonality. e.g a correlogram, box plot, line plot etc.
2. **Run Summary Statistics:-** Dividing the data into intervals and checking for clear or significant differences in the summary statistics of subdivided data sets. let's take a look at an example below

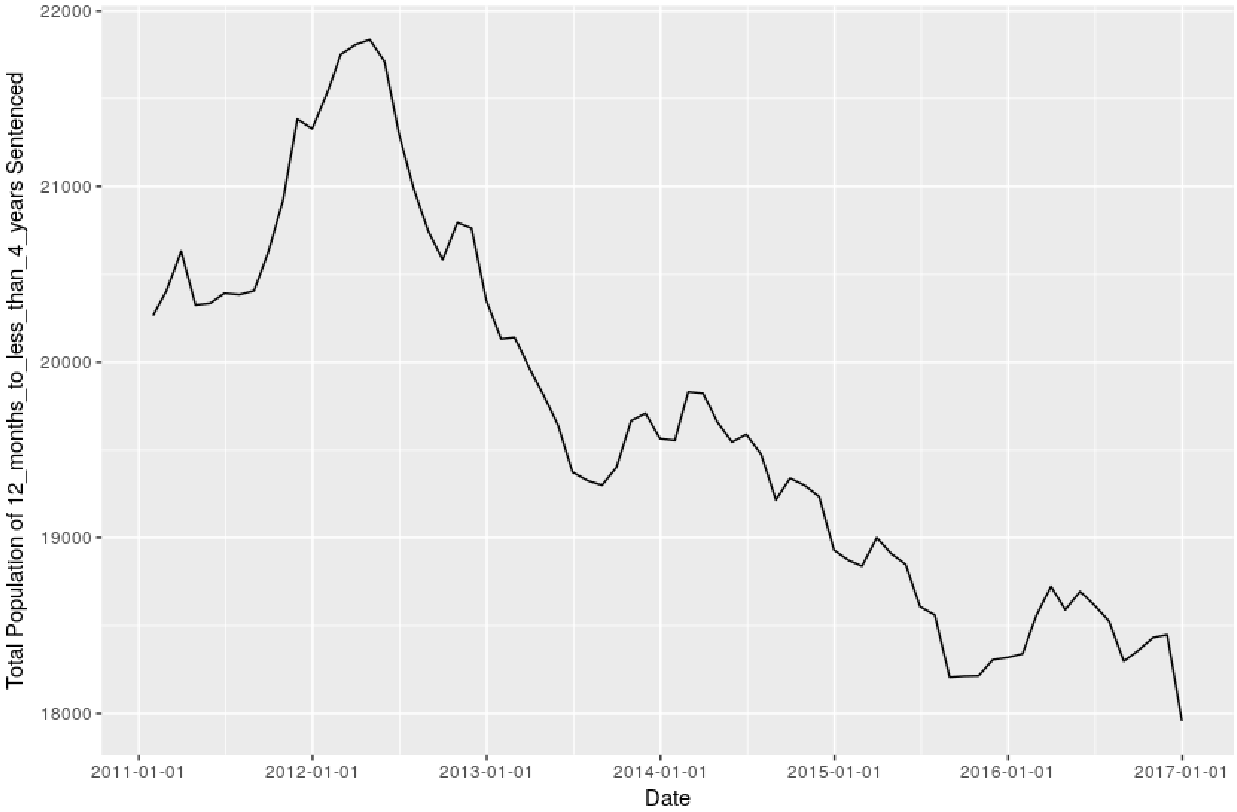


Figure 3.1: Sentenced length of 12 months to less than 4 years

From Figure 3.1, it is observed that there is a downward trend as time goes by, and we can also observe seasonality of similar shape in each year.

3.5 How to Cater for Stationarity

If we discover that a time series data is not stationary, we can often transform into stationarity by applying one of the following techniques.

1. **Creating a new series:-** This is called differencing the data set. Given a series X_t we can create a new series. i.e
$$X_t' = X_t - X_{t-1} \tag{3.5.1}$$
In this method, the differenced data will contain one less point than the original data set. Although this can be done more than once, one or two differencing is often sufficient.

2. **The Residual plot:**- Modeling the residuals from the fitted plot is helpful in removing long term trend, a simple fit, i.e the straight line is often used.

3. **Series Transformation :**- To account for non-constant variance, taking the logarithm or the square root of the data can help stabilize the variance. For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This introduced constant can then be subtracted from the model to obtain predicted (i.e. the fitted) values and forecasts for future points [20].

3.6 Models in Time Series

1. **AutoRegressive Models AR(p):**- An autoregressive model can be viewed as a representation of a type of random process, where the output variable depends linearly on its own previous values. Due to the randomness, predictions might not be 100 percent accurate, although most times, the process gets "close enough" for it to be useful in most scenarios [2]. The value for **"p"** is known as the order. e.g, an AR(1) would be called a *first-order autoregressive process*. The outcome variable in the AR(1) process at some point in time t is related only to time periods that are one period apart i.e. the value of the variable at t-1. AR(2) or AR(3) process would be related to data two or three periods apart. This is given as

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + e_t \tag{3.6.1}$$

Where:

- $X_{t-1}, X_{t-2}, \cdots X_{t-p}$ are the past series values also known as lags;
- e_t is white noise (i.e. randomness);
- $\alpha_1, \alpha_2,$ are weights measuring the influence of these preceding values on the value X_t, and c is a constant.

2. **Moving Average Models MA(q):**- This is the approach for modeling univariate time series. This model specifies that the output variable depends linearly on the current and various past values, as in the case of AR(P), the value for **"q"** is known as the order. e.g

$$X_t = c + \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \cdots + \alpha_q e_{t-q} + e_t \tag{3.6.2}$$

Where at a particular time t, the process takes the value X_t which is modeled as the current point in the white noise e_t and a constant a_q multiplied by the previous noise value e_{t-q}. Note that noise points are assumed to be independent and identically distributed i.e it follows a Gaussian Normal distribution with mean 0 and variance σ^2. The basic difference between AR(p) model and MA(p) model is that the covariance between $X(t)$ and $X(t-n)$ is zero for MA models. But that of the AR model gradually declines as n increases, meaning that the MA model does not predict the future using the past, but rather it uses the error from the past. However, the AR model uses past forecasts to predict future values.

3. **AutoRegressive Moving Average models ARMA(p,q):**- The combination of both AR and MA model is known as the ARMA model. Since an ARMA model is a stationary model, but working

with a model that isn't stationary, one can achieve stationarity by taking a series of differences. This difference changes an **ARMA** model to the **ARIMA** model. The **"I"** in the ARIMA model stands for integration or differencing; it measures the amount of nonseasonal differences that are needed to achieve stationarity. If no differencing is involved in the model, then it becomes simply an ARMA.

However, ARIMA(p,d,q) also known as the Box-Jenkins model is an ARMA(p,q) model (3.6.3) which has been differenced d times in order to fit or achieve stationarity. The choice of p,d, and q can be selected with a wide range of methods, including AIC, BIC, and empirical autocorrelations. [3]

$$X_t = c + e_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i e_{t-i} \tag{3.6.3}$$

Where:

- α_i are the autoregressive model's parameters;
- β_i are the moving average model's parameters;
- c is a constant;
- e_t is the error terms (white noise).

3.7 Correlation Function

There are two major types of Correlation Functions namely:

(a) **AutoCorrelation Function**:- In time series, the ACF explains the degree of similarity between the values of the same variables over successive time intervals. This is diagnosed by using a correlogram (ACF plot). If the existence of autocorrelation is detected in a model, it implies that the model is misspecified, i.e something is wrong with the model, maybe there are some key variable or variables that are missing from the model [7].

The autocorrelation at time lag zero, can be written as ACF(0)=1 meaning that the data is perfectly correlated with itself, ACF(1)=0.9 implies that the correlation between a point and the next point is 0.9, ACF(2)= 4 suggests that the correlation between a point and a point with 2 time lags or time steps ahead is 0.4. Hence, the ACF tells you how correlated points are with each other, based on how many time steps they are separated by [5]. This is a common tool used for identifying the order of an Moving Average model.

(b) **PartialAutoCorrelation Function**:- Partial autocorrelation plots are a commonly used tool for identifying the order of an autoregressive model. They are basically used to measure the degree of association between Y_t and Y_{t-p}, when the effects at the other time lags (i.e the intermediaries) eg $1, 2, 3, \cdots (p-1)$ are all removed. Since there is no intermediaries dependence, In principal, PACF and ACF at lag 1 are equal. The ACF for a stationary time series Y_t is equal to the autocorrelation, so ACF(1) $=$ Corr(Y_t, Y_{t-1}) $=$ PACF(1).

3.8 Diagnostic Checking

To avoid over fitting, we need to subject the model to a series of statistical tests, this is to ensure that this model adequately describes the time series under consideration; all of this process is known as Diagnostic Checking. For this study, we shall be considering two types of diagnostic checking.

1. **Akaike Information Criterion:-** Using the values of AIC/BIC/SBIC, the model with the lowest value of the above criterion is chosen as the best model, as the AIC compares the quality of a set of statistical models to each other.

$$AIC = -2(loglikelihood) + 2K$$

- K is the number of model parameters i.e the variables in the model as well as the intercept.

- Log-likelihood usually obtained from the statistical output is a measure of model fit. The higher the number, the better the fit.

2. **Ljung-Box Test:-** This hypothesis tests whether a set of autocorrelations of a fitted time series model differs significantly from zero, i.e whether or not the errors are independently identical. It essentially tests for lack of fit, such that if the autocorrelations of the residuals are very small, then we conclude that the model doesn't show significant lack of fit'.

Hypothesis:-

H0:- The correlations between the population series values are zero. i.e the model is fine.

H1:- The correlations between the population series values are not zero. i.e the model is not fine

$$P(m) = n(n+2)\sum_{i=1}^{m} \frac{R_i^2}{n-i} \tag{3.8.1}$$

R_i Is the sample autocorrelations at lag i, m is the number of lags under the test, n is the length of the time series sample.

Decision Rule:-

For a chi-squared distribution with h degrees of freedom at the $100(1-\alpha)th$ percentile. We reject $H0$ if

$$P(m) > \mathbb{X}^2_{\alpha,h}$$

Concluding that the correlations between the population series values are zero. i.e the model is fine and does not show a lack of fit. Also, a significant p-value in this test rejects the null hypothesis that the time series isn't autocorrelated.
For the purpose of this study, we will be using a statistical software "R", to implement the Ljung-Box Test with the Box.test function.

3. **Forecast Accuracy:-** After getting the right model, the next thing is to make a forecast, but one way to evaluate the forecast's accuracy is by comparing the forecast to the past actual data, thus, it is necessary to assume that a forecast will be as accurate as it has been in the past, and that future accuracy of a forecast can be guaranteed. This can be done by calculating 3 different measures.

(a) **Mean Absolute Error (MAE):-**"Error" which does not necessarily mean mistake but the unpredicted part of an observation. MAE is defined as the absolute value of the difference between the forecasted value and the actual value. **MAE** shows how big the error we can expect from the forecast on average. For a time series analysis the (MAE) can also be seen as the **Forecast Error** i.e the difference between the actual and the predicted or forecast value of a time series [8]. Hence, the smaller the MEA, the better the prediction. MAE is given as:-

$$\text{MAE} = e(t) = \sum_{i=1}^{n} |y_i - x_i|$$

Where:

- $e(t)$ is the forecast error;
- x_i is the actual value;
- y_i is the forecast value.

Sometimes it is always difficult to tell how big an error is, hence, it is better to find the mean absolute error in percentage terms.

(b) **Mean Absolute Percentage Error (MAPE):-** This is the percentage form of the (MAE). it is given as:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - P_t}{A_t} \right|$$

Where:

- A_t is the actual error;
- P_t is the predicted or forecasted value.

Although (MAPE) allows us to compare forecasts of different series in different scales unlike the MA, but it is some-worth limited in the sense that, it is undefined when we have one of the actual value to be zero, also it may not really account for large errors, so to adjust for large errors we consider the nest text.

(c) **Root Mean Square Error (RMSE):-** This is also known as the standard deviation of the residuals or prediction errors. Since residual is a measure of how far data points are from the regression line, hence RMSE measures how these residuals spread out. i.e it enables us to know how close the data is around the line of best fit.

$$\text{RMSD}(\hat{\theta}) = \sqrt{\text{MSE}(\hat{\theta})} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}$$

Where $\hat{y}_t$ is the predicted values for time t of a regression dependent variable y_t over T different prediction.

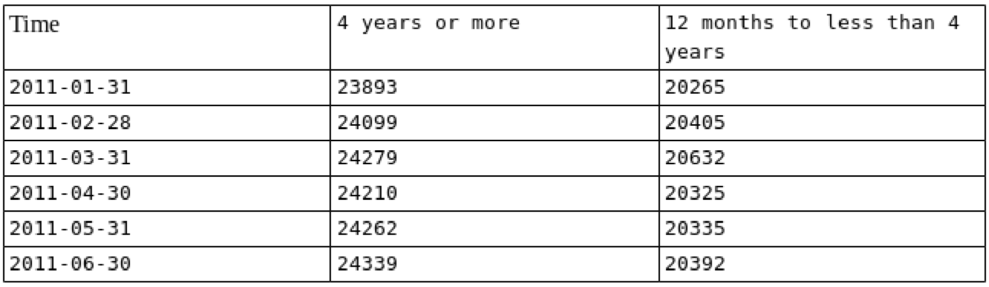
4. Analysis and Discussion

4.1 Data presentation

The focus of this chapter is to conduct a descriptive analysis as well as a predictive analysis. To start with, we need to get the summary of the data and also perform some data visualization after which we begin with a simple model (Linear Regression Model) after which we move to the (Time Series), and finally, (Stock Flow model). and then make forecast or predictions.

4.2 Date preparation, Exploration and Visualization

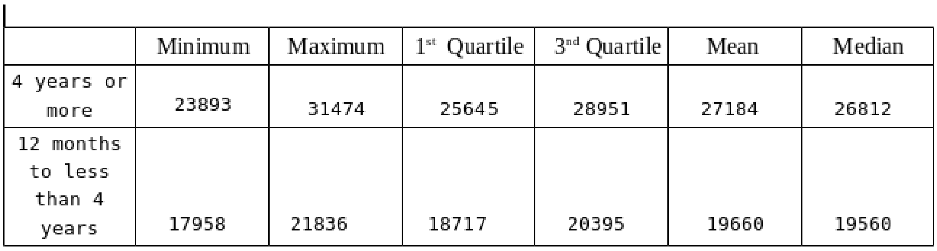
This data was collected online at [14] from England and Wales, it consists of the prison population by type of custody, age group, and sex and it was collected monthly, beginning from 31 January 2011 to 31 December 2016. The data is categorized into different sections; (1) the remand and (2) the sentenced population. The sentenced population is further subdivided into different categories, for the purpose of this study, we shall be considering the category of sentence length of **4 years or more (excluding indeterminate sentences)**.



Time	4 years or more	12 months to less than 4 years
2011-01-31	23893	20265
2011-02-28	24099	20405
2011-03-31	24279	20632
2011-04-30	24210	20325
2011-05-31	24262	20335
2011-06-30	24339	20392

Figure 4.1: The first 5 head of the data set

Figure 4.1 displays the first 5 rows of the data set. from the table, it is observed that the data were collected monthly.



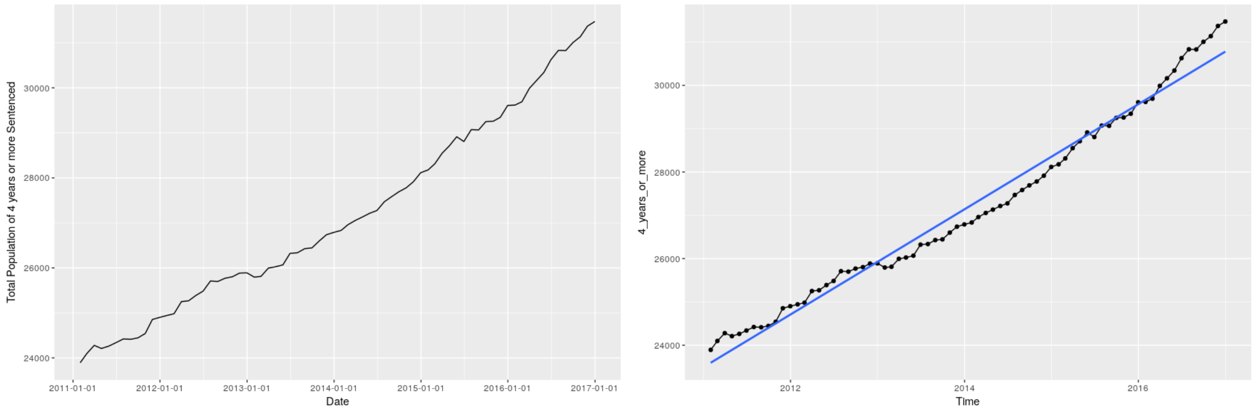
	Minimum	Maximum	1^{st} Quartile	3^{rd} Quartile	Mean	Median
4 years or more	23893	31474	25645	28951	27184	26812
12 months to less than 4 years	17958	21836	18717	20395	19660	19560

Figure 4.2: The Descriptive Statistics

Figure 4.2 gives the summary of the data set ranging from its minimum to it's maximum.

4.3 Regression Analysis

Linear Regression helps to model the relationship between two variables. i.e explanatory variable, and the other is a dependent or response variable.



(a) Sentenced length of 4 years or more

(b) The model and the line of best fit

Figure 4.3: The Regression model

Figure 4.3a shows a time plot of prisons population for the Sentenced length of 4 years or more, while Figure 4.3b shows the fitted model and the line of best fit. From the plot, it is observed that most of the point deviates from the line of best fit.

4.3.1 Decomposition. This is a way of breaking or simplifying time series data into systematic (level, trend, and seasonality) and unsystematic components (noise), for a better understanding. From Figure 4.3a it is observed that the changes do not change over time, i.e the trend is linear and the linear seasonality has the same frequency and amplitude over time. Hence Figure 4.4 shows an additive decomposition.

$$y(t) = Level + Trend + Seasonality + Noise$$

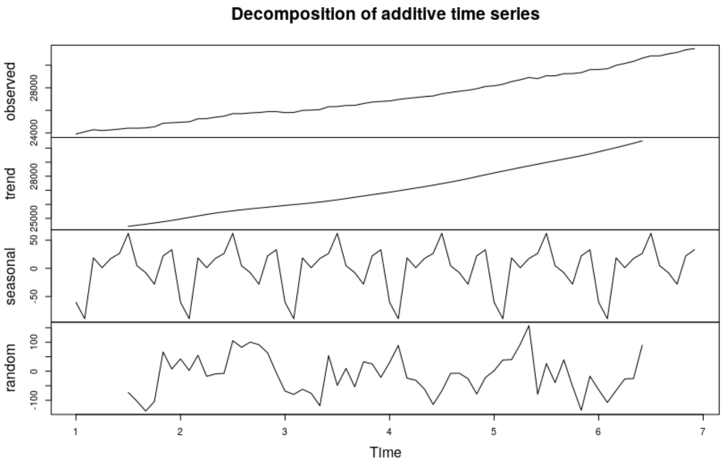


Figure 4.4: Time Series Decomposition of Sentence length of 4 years or more

From the plot of Figure 4.4, we can see the evidence of trend, seasonality, and noise. (All these where not visible by simply looking at Figure 4.3a) The trend shows a linear increase with time, and the seasonality shows a sharp decrease in the prison population at the beginning of each year of 2011 - 2016, which is being represented by 1 - 6, but a sharp increase at the middle of each year. The presence of seasonality and trend shows that the data is not stationary.

Due to the presence of seasonality and trend, there is a need to introduce the smoothing spline to avoid the issue of over-fitting by using regularized regression. This involves minimizing a criterion that includes both a penalty for the least squares error and roughness penalty.

4.3.2 Spline Functions. These are functions used to approximate nonlinear functions, this is achieved by dividing the domain of the variable into contiguous intervals, after which we fit separate polynomials within each range. Knots are placed at the interval endpoints where the polynomials are joined. Let the interval be $[x, y]$

$$x = \xi_0 < \xi_1 \cdots \xi_N < \xi_{N+1} = y \tag{4.3.1}$$

where the N knots $\xi_1 \cdots \xi_N$, for $n = 1, 2 \cdots, N$ are known as the inner knots. A spline of degree p has the following properties.

(a) f is a piecewise polynomial such that on each interval $[\xi_1 \cdots \xi_N]$ is a polynomial of degree q

(b) f is $p - 1$ times continuously differentiable at the knots.

Generally, a spline of degree 1 is a piecewise linear function where the straight line segment connects at the knots. A spline of degree 2 is a piecewise quadratic curve with continuous first derivative at the knots. while a spline of degree 3 is a piecewise cubic function which has continuous first and second derivatives at the knots.

A linear model is of the form

$$Y_i = \alpha + \beta X_i + e_i$$

where $i = 1 \cdots n$ α, β are model parameters i.e the intercept and slope of the line and X_i is the explanatory or independent variable (Time), e_i is an additive or the unpredictable quantity, Y_i is the response or dependent variable (Prison population of sentence length 4 year or more).

Now, introducing the spline function of time t with a degree of freedom $(\lambda = 14)$in order to control for seasonality and trend we have

$$Y_i = \alpha + (t, \lambda) + e_i$$

From literature, λ is calculated as $3 * n - 1$, 3 to 7 degree of freedom per year has been justified as a balance to provide adequate control for seasonality and trends in time, while n is the number of calender years in the data (2011 - 2016). hence we have

$$3 \times 5 - 1 = 14$$

From A.1 (Appendix), it is observed that the Adjusted R-square is 0.9991, which means that 99.91 % of variation in the responds (prison population) variable is explained by the explanatory variable (Time)

4.4 Summary of the fitted model

The Total Sum Of Square is the difference between the observation and the mean. This is given as

$$TSS = \sum_i (y_i - \bar{y})^2 = S_{yy}$$

The Residual Sum Of Square is the difference between the observation and the fitted model. This is given as

$$RSS = \sum_i (y_i - \hat{y})^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

The Measure of model fit shows how much variation in the data our model has explained. This is given as

$$R^2 = 1 - \frac{RSS}{TSS}$$

For the simple linear model $R^2 = r^2$. hence, we show that $R^2 = r^2$, where R^2 is the coefficient of determination and r is the sample correlation coefficient.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx}.S_{yy}} = r^2$$

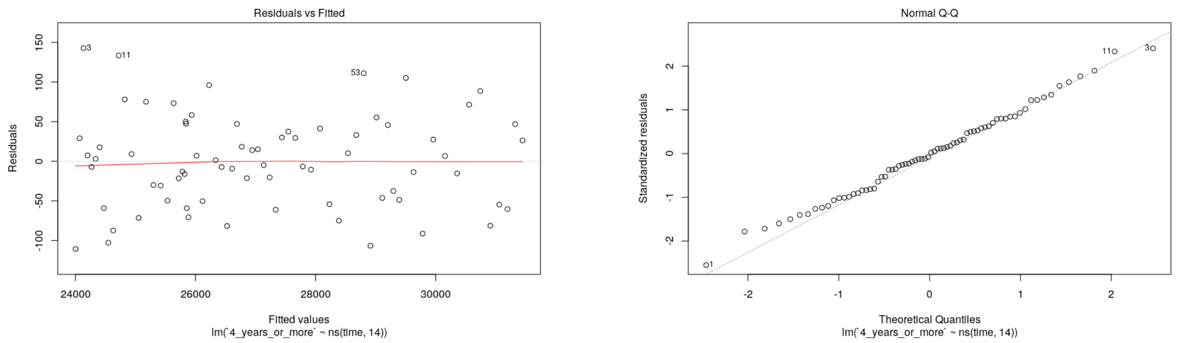
4.5 Assumptions of a Linear Model

To make sure that the model and techniques we have used are the appropriate ones, we have to apply the general methods for checking regression models. And this is known as the model assumptions [1].

1. The error term has a population mean of zero, meaning that for a model to be unbiased, the average value of the error term must equal zero.
2. The error term has a constant variance (Homoscedasticity), that is, it must not start out little and then start spreading of vice versa
3. The errors are independent, which implies that the knowledge of the error attached to one observation does not give us any information about the error attached to another.
4. The error term is normally distributed, This allows us to describe the variation in the model's parameter estimates, which in turn enable us to make inference about the population we are considering.

4.6 Residuals plot

Once a regression model has been fitted, there is a need to examine the residuals, that is, the deviations from the fitted line to the observed values. With this we can investigate the validity of the assumption that a linear relationship exists between the response variable and the explanatory variable. The residuals plot reveals the true relationship between the variables. It amplifies the presence of outliers, as well as the possibility of a non-linear relationship among the variables.



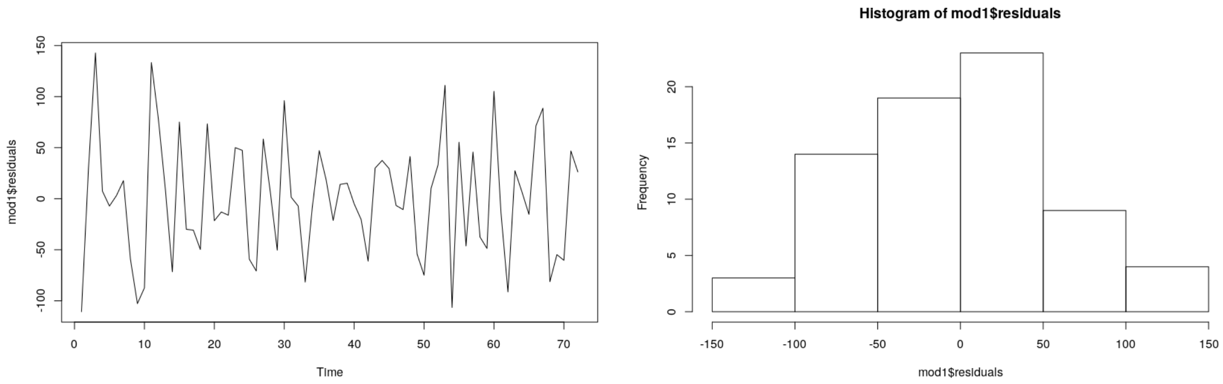
(a) The Residuals Plot

(b) The normal Q-Q Plot

Figure 4.5: The Residuals alongside the Normal QQ- Plot for the linear regression model

From the residual plot in Figure 4.5a, it is observed that it the first two assumptions are met, i.e the population of the error term is almost at zero (this can be seen from the red line drawn horizontally). and also the error term does have a constant variance, (Homoscedasticity). Also from 4.5b, the normal Q-Q plot also meets the fourth assumption which states that the error term is normally distributed. Although we can still see some data points falling off the line especially at the left tail of the plot.

4.6.1 Forecasting with the Regression Model. Figure 4.14 shows the forecast for the next 27 month, alongside the validating data set. To investigate if the forecast errors are normally distributed with mean zero and constant variance, there is a need to make a time plot of the residual and the fitted model as well as a histogram of the forecast errors.



(a) The Residuals Plot

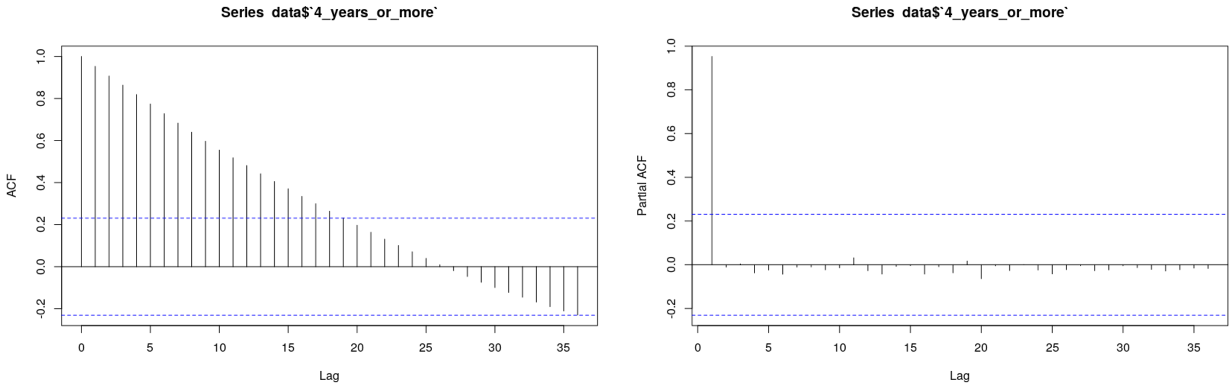
(b) The normal Q-Q Plot

Figure 4.6: The Residuals alongside the Normal QQ- Plot

In Figure 4.6a, the time plot of the forecast errors shows that the variance of the forecast errors seems to be roughly constant over time although there seems slightly higher variance at the beginning and towards the tail end i.e from 2011 to about middle of 2012 and then it became a bit constant but started rising from 2015 up till the end of 2016. Figure 4.6b shows that the histogram of the forecast errors are roughly normally distributed and the mean seems to be close to zero. Hence the above model seems to provide an adequate predictive model for the prisons population of sentence length of 4 years or more.

4.7 Time Series Analysis

As discussed in chapter 3, let's consider another model, a time series analysis is the analysis of a series of observations collected over time. Since the data had been decomposed in figure 4.4 we can also confirm the presence of seasonality and trend by also plotting its ACF and PACF.



(a) The ACF Plot

(b) The PACF Plot

Figure 4.7: The Correlation Functions

Figure 4.7a shows the ACF plot of the prison population for sentence length of 4 years or more for lag 36, i.e 3 years, the plot shows the presence of seasonality after 2 years, this is noticed at lag 24. The presence of the trend is also visible as there is a downward decrease in the lags. The PACF in 4.7 shows that there is a partial autocorrelation at lag 1, i.e a sharp cut-off at lag 1

4.8 Data Differencing

Since the data is nonstationary, we proceed to differencing the data. i.e creating a new series. where

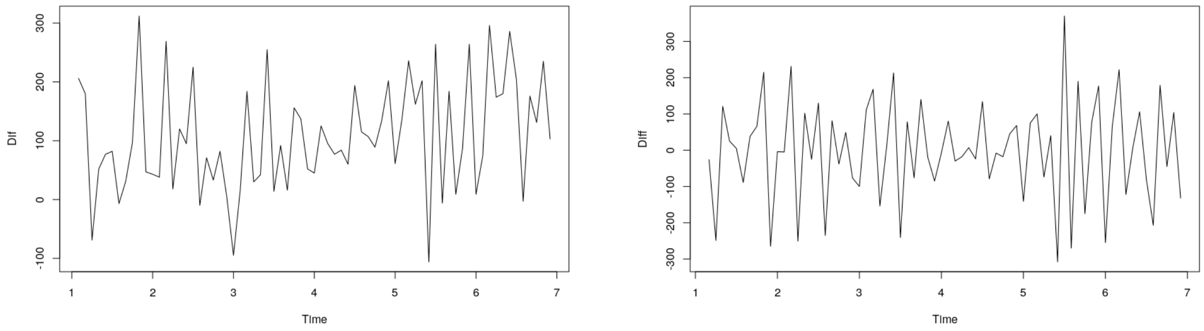
$$Y_t' = Y_t - Y_{t-1}$$

Since the first differencing does not give a stationary data, we proceed to the 2nd differencing, we have

$$Y_t'' = Y_t' - Y_{t-1}'$$

$$Y'' = Y_t - Y_{t-1} - Y_{t-1} - Y_{t-2}$$

$$Y'' = Y_t - 2Y_{t-1} - Y_{t-2}$$

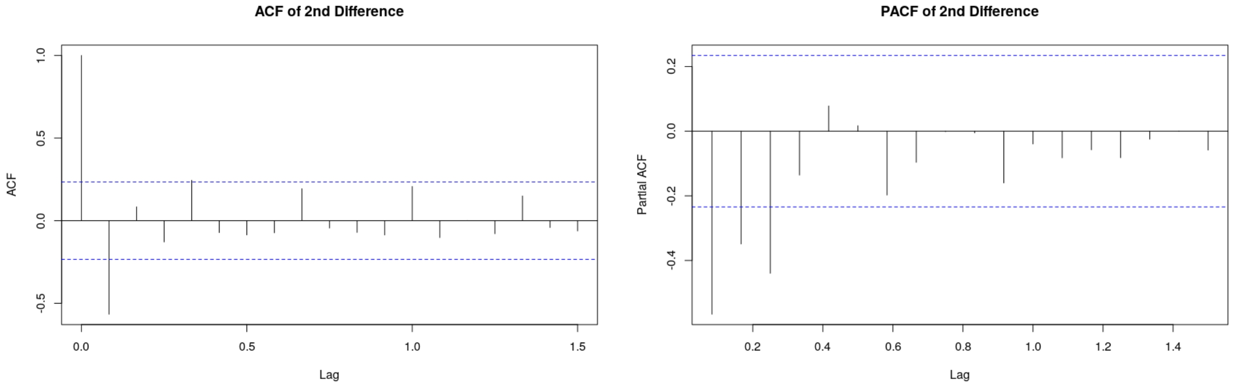


(a) The First Difference

(b) The Second Difference

Figure 4.8: The First and Second differencing of the data

Figure 4.8a is the 1st difference of the data set, the mean is still not close to zero, i.e, not stationary. Figure 4.8b shows the plot of the 2rd difference of the data set. From the plot, the time series of 2rd differences appears to be stationary in mean and variance, hence an ARIMA(p,2,q) model seems to provide an adequate predictive model for the prisons population of sentence length of 4 years or more.



(a) ACF of 2rd Difference

(b) PACF of 2rd Difference

Figure 4.9: The ACF and PACF of the 2nd differencing

4.9 Model Identification

Having differenced the data and plotted the ACF and PACF, how then can we identify the model to use? Just as we discussed in the previous chapter the ACF helps to know the number of lag "p" while PACF helps to identify the number of lag "q" From the plot above, **p** = 0,1,4 and **q**= 1,2,3 since we have differenced twice, then **d** = 2

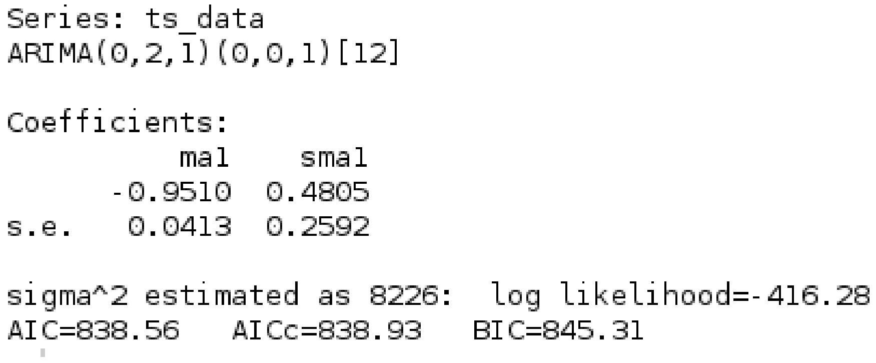
4.9.1 Hints to Identifying a model.

1. if the PACF of the differenced series displays sharp cut-off while the ACF decays slowly, it implies that the autocorrelation pattern is better explained more by adding *AR* terms.
2. if the PACF of the differenced series displays sharp cut-off and or the first lag is **positive**, then consider adding an *AR* term to the model
3. if the ACF of the differenced series displays sharp cut-off and or the first lag is **negative**, then consider adding an *MA* term to the model [4].

Looking at the differenced ACF and PACF plot alongside the hint provided above, we can see that the first lag of the differenced ACF is negative and there is also a sharp cut off between lag 1,2 3 and lag 4. Hence, we consider adding an MA model. Also, the differenced PACF shows a sharp cut off between lag 2,3 and lag 4. Hence, we also consider adding an AR model.

4.10 Model Estimation

Having identified the model to use, how can we then estimate the best model? From **p** and **q** identified earlier, we shall consider **ARIMA (021), (022), (023), (121), (122), (123)**. Since we have so many models to work with, we shall introduce the function **auto.arima()** in R. This function takes into account the AIC and BIC values to determine the best combination of parameters. It also gives the model with the important parameters.



```
Series: ts_data
ARIMA(0,2,1)(0,0,1)[12]

Coefficients:
          ma1    sma1
      -0.9510  0.4805
s.e.   0.0413  0.2592

sigma^2 estimated as 8226:  log likelihood=-416.28
AIC=838.56   AICc=838.93   BIC=845.31
```

Figure 4.10: The auto.arima model

4.11 Mathematical Formulation of The ARIMA Model

Starting with the backward shift operator, let's consider

$$BZ_t = Z_{t-1} \tag{4.11.1}$$

B causes the observation that it multiplies to be shifted backward in time by 1 period. i.e, for any time series Y and any period t, increasing the power of B by 1 we have;

$$B^2Z_t = B(BZ_t) = B(Z_{t-1}) = Z_{t-2} \tag{4.11.2}$$

Hence in general, for any integer n, it has the effect of shifting an observation backward by n periods.

$$B^n Z_t = Z_{t-n}$$

Recall from equation 3.5.1, let's consider a simple example of the first-difference operation. Suppose that z is the first difference of Z. Then, for any t

$$z_t = Z_t - Z_{t-1}$$

$$= Z_t - BZ_t$$

$$= (1-B)Z_t$$

Since the *first differenced* series z is obtained by multiplying a factor of $(1-B)$ to the original series Z

Hence multiplying a factor of $(1-B)^2$ to the original series Z we'll obtain the *second differenced* series.

$$(1-B)((1-B)Z_t) = (1-B)^2 Z_t$$

Therefore, generally multiplying by a factor of $(1-B)^d$ gives the *dth* difference of Z

4.11.1 The Seasonal ARIMA Model. From figure 4.10, the output of the auto.arima model indicates a SARIMA Model, i.e ARIMA (0.2.1)(0,0,1)[12]. The seasonal ARIMA model incorporates both the non-seasonal and seasonal factors in a multiplicative model. This is given as

ARIMA $(p, d, q) \times (P, D, Q)S$

where **p** is the nonseasonal AR order, **d** is the non-seasonal differencing, **q** is the non-seasonal MA order, **P** is the seasonal AR order, **D** is the seasonal differencing, **Q** is the seasonal MA order, and **S** is the time span of repeating seasonal pattern [6].

The general multiplicative of a SARIMA Model is given as;

$$\alpha_P(B^s)\beta_p(B)(1-B)^d(1-B^s)^D \; Z_t \;\; = \;\; \theta_q(B)\gamma_Q(B^s)e_t \tag{4.11.3}$$

From the backward shift in equation 4.11.1, equation 4.11.3 can be further simplified as follows

$$\alpha_P(B^s) = 1 - \alpha_1(B^s) - \alpha_2(B^{2s}) - \cdots - \alpha_P(B^{Ps}) \tag{4.11.4}$$

$$\beta_p(B) = 1 - \beta_1(B) - \beta_2(B^2) - \cdots - \beta_p(B^p) \tag{4.11.5}$$

$$\theta_q(B) = 1 - \theta_1(B) - \theta_2(B^2) - \cdots - \theta_q(B^q) \tag{4.11.6}$$

$$\gamma_Q(B^s) = 1 - \gamma_1(B^s) - \gamma_2(B^{2s}) - \cdots - \gamma_Q(B^{Qs}) \tag{4.11.7}$$

Equation 4.11.5 and 4.11.6 are known as the nonseasonal components of the model, while 4.11.4 and 4.11.7 are the seasonal components of the model.

Hence from the output in figure 4.10, SARIMA Model (0.2.1)(0,0,1)[12] can be calculated as follows; where **p**$=0$, **d** $=2$, **q** $=1$, **P** $=0$, **D** $=0$, **Q** $=1$ and s$=12$

Substituting these values into equation 4.11.3 we have

$$\alpha_0(B^{12})\beta_0(B)(1-B)^2(1-B^{12})^0 Z_t = \theta_1(B)\gamma_1(B^{12})e_t \tag{4.11.8}$$

But $\quad \alpha_0(B^{12}) = 1; \qquad \beta_0(B) = 1; \qquad (1-B^{12})^0 = 1$

Hence we are left with

$$(1-B)^2Z_t = (1-\theta_1 B)(1-\gamma_1 B^{12})e_t \tag{4.11.9}$$

expanding equation 4.11.9, we have

$$(1-2B+B^2)Z_t = (1-\gamma_1 B^{12} - \theta_1 B - \theta_1\gamma_1 B^{13})e_t \tag{4.11.10}$$

Recalling equation 4.11.1, and substituting in the above equation, we have;

$$Z_t = 2Z_{t-1} - Z_{t-2} + e_t - \gamma_1 e_{t-12} - \theta_1 e_{t-1} - \theta_1\gamma_1 e_{t-13} \tag{4.11.11}$$

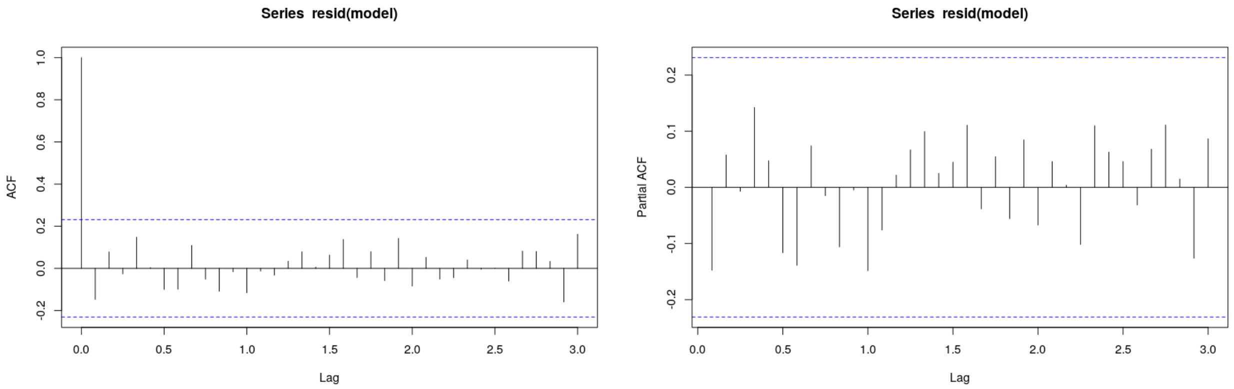
Hence, equation 4.11.11 is the required model

4.12 Diagnostic Checking

1. **Akaike Information Criterion:-** Since the function **auto.arima()** takes into account the AIC and BIC values to determine the best combination of parameters, hence the AIC is **838.93**

2. **Ljung-Box Test:-** Using a statistical software "R", to implement the Ljung-Box Test with the Box.test function. figure A.2 shows that the **p-value** which is **0.8562** is highly significant, Therefore we do not reject H0, hence, we conclude that the correlations between the population series values are zero. i.e the model is fine and does not show a lack of fit.

3. **ACF and PACF plots of Residual**



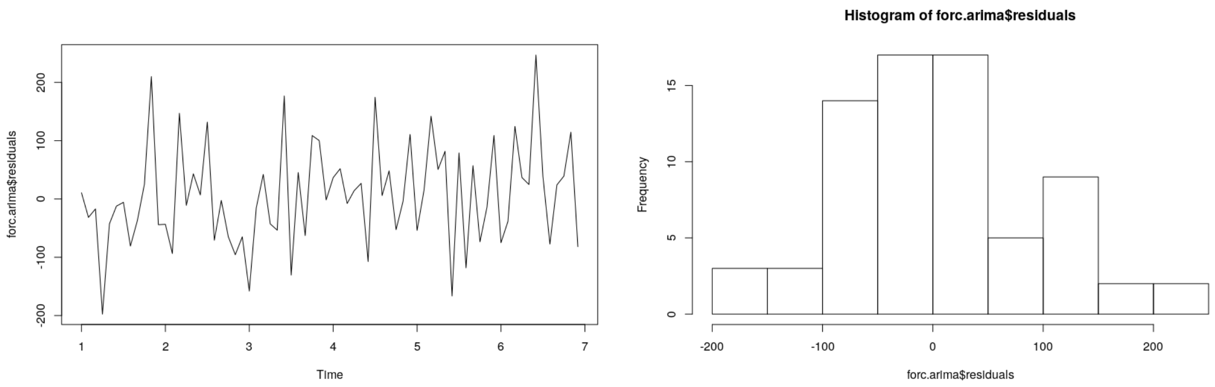
(a) ACF of Model Residuals

(b) PACF of Model Residuals

Figure 4.11: The Model Residuals

From the above ACF and PACF plots of the residual, it is observed that all the lags lie within the 95% confidence interval in both plots, and also the PACF shows no autocorrelation.

4. **Time plot and histogram of Residual**



(a) Residuals of the fitted model

(b) Histogram of the residuals

Figure 4.12: The Model Residuals

From the time plot of the residuals in figure 4.12a, we observed that the randomness was within mean of zero with a constant variance, the histogram in figure 4.12b also shows that the residuals are normally distributed with mean of zero.

4.13 Predictive Analysis

Having conducted the diagnostic checking and seen that the chosen model does not violate any of the rules, predictions can be done.

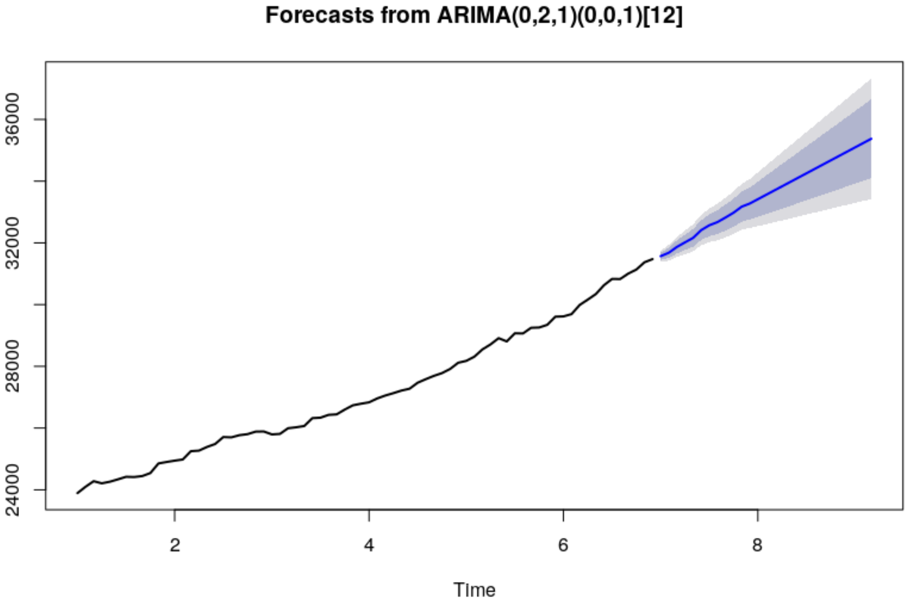


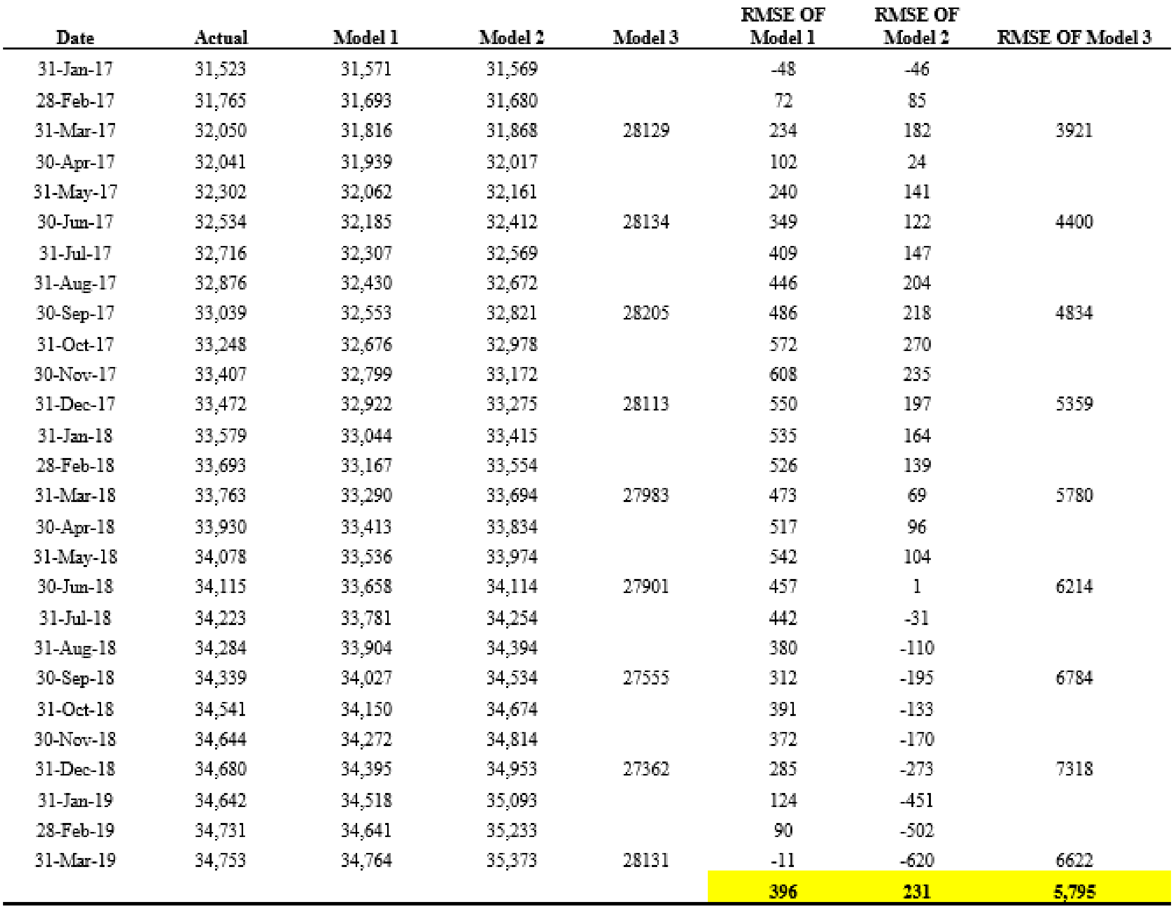
Figure 4.13: 27 Months Forecast of he Prison Population using the SARIMA Model

Figure 4.13 is the 27 months forecast of the prison population using the ARIMA (0,2,1)(0,0,1)[12] model with 80% and 95% confidence interval. Using the prison population starting from January 2011 to December 2016, and the forecasting for January 2017 to March 2019. 2,4,6 represents 2012, 2014, and 2016 respectively.The thick blue line is the actual forecast, while the dark ash region represents the 95% confidence interval, and the fainted ash region represents the 80% confidence interval.

4.14 Forecast Accuracy

It is one thing to make predictions, but it is another thing for the predictions to be accurate. Below are some of the ways to test if the forecast is accurate.

Validation Test:- This is checking if the forecasted values also tallies with a set aside observation or data which is not part of the training set.



Date	Actual	Model 1	Model 2	Model 3	RMSE OF Model 1	RMSE OF Model 2	RMSE OF Model 3
31-Jan-17	31,523	31,571	31,569		-48	-46	
28-Feb-17	31,765	31,693	31,680		72	85	
31-Mar-17	32,050	31,816	31,868	28129	234	182	3921
30-Apr-17	32,041	31,939	32,017		102	24	
31-May-17	32,302	32,062	32,161		240	141	
30-Jun-17	32,534	32,185	32,412	28134	349	122	4400
31-Jul-17	32,716	32,307	32,569		409	147	
31-Aug-17	32,876	32,430	32,672		446	204	
30-Sep-17	33,039	32,553	32,821	28205	486	218	4834
31-Oct-17	33,248	32,676	32,978		572	270	
30-Nov-17	33,407	32,799	33,172		608	235	
31-Dec-17	33,472	32,922	33,275	28113	550	197	5359
31-Jan-18	33,579	33,044	33,415		535	164	
28-Feb-18	33,693	33,167	33,554		526	139	
31-Mar-18	33,763	33,290	33,694	27983	473	69	5780
30-Apr-18	33,930	33,413	33,834		517	96	
31-May-18	34,078	33,536	33,974		542	104	
30-Jun-18	34,115	33,658	34,114	27901	457	1	6214
31-Jul-18	34,223	33,781	34,254		442	-31	
31-Aug-18	34,284	33,904	34,394		380	-110	
30-Sep-18	34,339	34,027	34,534	27555	312	-195	6784
31-Oct-18	34,541	34,150	34,674		391	-133	
30-Nov-18	34,644	34,272	34,814		372	-170	
31-Dec-18	34,680	34,395	34,953	27362	285	-273	7318
31-Jan-19	34,642	34,518	35,093		124	-451	
28-Feb-19	34,731	34,641	35,233		90	-502	
31-Mar-19	34,753	34,764	35,373	28131	-11	-620	6622
					396	231	5,795

Figure 4.14: Prison Population Forecast for 25 Months

Figure 4.14 shows the forecast for 27 more months, the Actual is the validation set. Model 1 is the forecast from the Regression Model with RMSE of **396**, Model 2 is the forecast from the Seasonal Arima Model with RMSE of **231**, and Model 3 is the forecast from the stock flow model with RMSE of **5795**. we observed that Model 2 which is the SARIMA model performed better than the others.

4.15 The Stock Flow Model

The two models studied above (Regression and Arima) predicted the prisons population without taking into account how the offenders are being admitted and how they are being released. To cater for that, there is a need for a stock-flow model figure 1.3 describes this.

This is given as;

$$P_{t+1} = IP_t + P_t - OP_t \tag{4.15.1}$$

Where at P_{t+1} is the new closing population stock estimated by adding the in-flow of offenders IP from time t to $t+1$ with the initial or starting prison population stock P at time t and subtracting the out-flow population OP at time t to $t+1$, and the result obtained (closing stock) becomes the (starting stock) for the next prison population. Example

Let's consider the prison population by type of custody, age group and sex for sentence length of 4 years or more(excluding indeterminate).

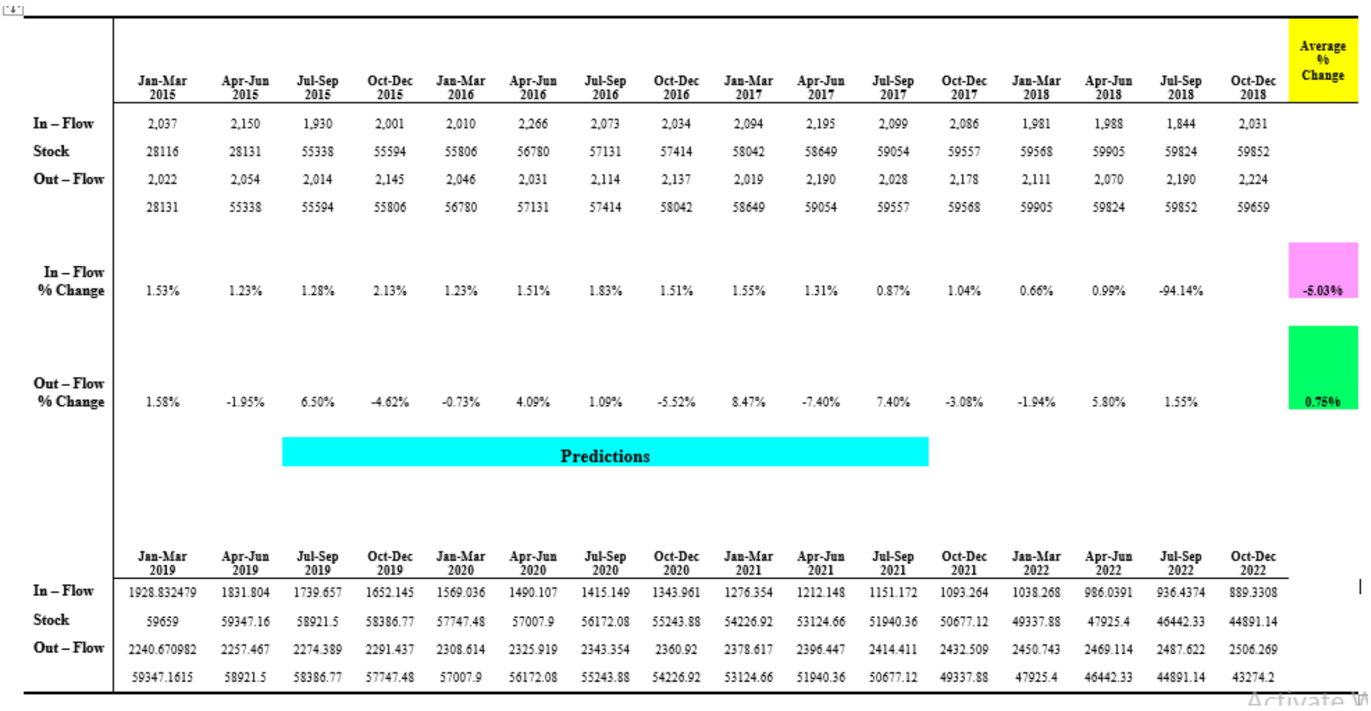
If the prison admission (**in-flow**) for October - December 2017 is 2086, and the prison population (**starting stock**) as at 30 September 2017 is 33039 and 2178 (**out-flow**) were released as at October - December 2017, what would be the closing stock for the month of October - December?

From equation 4.15.1 we have;

$$P_{t+1} = 2086 + 33039 - 2178 \tag{4.15.2}$$

$$NP_{t+1} = 32947$$

Using this example, we shall predict for more quarters by considering the percentage change (increase or decrease) across each quarter.



	Jan-Mar 2015	Apr-Jun 2015	Jul-Sep 2015	Oct-Dec 2015	Jan-Mar 2016	Apr-Jun 2016	Jul-Sep 2016	Oct-Dec 2016	Jan-Mar 2017	Apr-Jun 2017	Jul-Sep 2017	Oct-Dec 2017	Jan-Mar 2018	Apr-Jun 2018	Jul-Sep 2018	Oct-Dec 2018	Average % Change
In – Flow	2,037	2,150	1,930	2,001	2,010	2,266	2,073	2,034	2,094	2,195	2,099	2,086	1,981	1,988	1,844	2,031	
Stock	28116	28131	55338	55594	55806	56780	57131	57414	58042	58649	59054	59557	59568	59905	59824	59852	
Out – Flow	2,022	2,054	2,014	2,145	2,046	2,031	2,114	2,137	2,019	2,190	2,028	2,178	2,111	2,070	2,190	2,224	
	28131	55338	55594	55806	56780	57131	57414	58042	58649	59054	59557	59568	59905	59824	59852	59659	
In – Flow % Change	1.53%	1.23%	1.28%	2.13%	1.23%	1.51%	1.83%	1.51%	1.55%	1.31%	0.87%	1.04%	0.66%	0.99%	-94.14%		-5.03%
Out – Flow % Change	1.58%	-1.95%	6.50%	-4.62%	-0.73%	4.09%	1.09%	-5.52%	8.47%	-7.40%	7.40%	-3.08%	-1.94%	5.80%	1.55%		0.75%

Predictions

	Jan-Mar 2019	Apr-Jun 2019	Jul-Sep 2019	Oct-Dec 2019	Jan-Mar 2020	Apr-Jun 2020	Jul-Sep 2020	Oct-Dec 2020	Jan-Mar 2021	Apr-Jun 2021	Jul-Sep 2021	Oct-Dec 2021	Jan-Mar 2022	Apr-Jun 2022	Jul-Sep 2022	Oct-Dec 2022
In – Flow	1928.832479	1831.804	1739.657	1652.145	1569.036	1490.107	1415.149	1343.961	1276.354	1212.148	1151.172	1093.264	1038.268	986.0391	936.4374	889.3308
Stock	59659	59347.16	58921.5	58386.77	57747.48	57007.9	56172.08	55243.88	54226.92	53124.66	51940.36	50677.12	49337.88	47925.4	46442.33	44891.14
Out – Flow	2240.670982	2257.467	2274.389	2291.437	2308.614	2325.919	2343.354	2360.92	2378.617	2396.447	2414.411	2432.509	2450.743	2469.114	2487.622	2506.269
	59347.1615	58921.5	58386.77	57747.48	57007.9	56172.08	55243.88	54226.92	53124.66	51940.36	50677.12	49337.88	47925.4	46442.33	44891.14	43274.2

Figure 4.15: The Stock Flow Model of The Prisons Population Recorded Quarterly

Figure 4.15 displays the stock flow model recorded quarterly from January 2015 to December 2018. and predictions were made for January 2019 to December 2022.

The in-flow, stock, and out-flow from Jan-March 2015 was calculated using the formula from the equation 4.15.1 Where $IP_t = 2037, \quad OP_t = 2022$ and $P_{t-1} = 28116$ which is the closing stock of prison population as at 31st December 2014. The result obtained in cell B5 now becomes the closing stock for Jan-Mar, but a starting stock for Apr-June, and this continues till Oct-Dec 2018.

Now, there is a need to know the percentage change in the in-flow from one quarter to the next, and this is calculated using the formula below;

$$(t \div t_{-1}) - 1 \qquad\qquad \text{for } IP \text{ at } OP$$

i.e the percentage change between Apr-Jun and Jan-mar is given as

$$\frac{IP_t}{IP_{t-1}} = \frac{2150}{2037} - 1 = 5.55\%$$

Meaning that there is a 5.55% increase in the population of offenders coming in from the 1st quarter to the 2nd quarter in the year 2015.

The cell labeled purple is the Average overall percentage of the offenders coming in from Jan 2015 to Dec 2018, which is 0.19%

This same procedure is also used in calculating the outflow.

$$\frac{OP_t}{OP_{t-1}} = \frac{2054}{2022} - 1 = 1.58\%$$

Meaning that there is a 1.58% increase in the population of prisoners being released from the 1st quarter to the 2nd quarter in the year 2015.

The cell labeled green is the Average overall percentage of the prison population being released from Jan 2015 to Dec 2018, which is 0.75%

Predictions

From the steps above, we can make predictions.

The inflow for Jan-Mar 2019 = $IP_{t-1}\times$ (1 + Average %IP)

= 2031(1+ 0.19%)

= 2034.762

The Outflow for Jan-Mar 2019 = $OP_{t-1}\times$ (1 + Average %OP)

= 2224(1+ 0.75%)

= 2240.671

Hence, the closing stock of Jan-Mar 2019 is calculated using equation 4.15.1

= 27156.09

Model Validation Comparing the estimated closing stock of each month with its actual values in figure 4.14. It is observed that the closing stock obtained is a bit lesser than the actual values. This might be as a result of some factors not considered in the model, i.e **Recidivism**

4.16 Recidivism

This is defined as the tendency of an ex-prisoner to get rearrested again after being released from prison. According to the National Statistics on Recidivism, it was recorded that in England and Wales as of 2005, 41.6% of adult ex-prisoners re-offended within a year of being released from custody. Between April 2010 and March 2011, 56,000 adult offenders were released from prison, but out of this number, around 26,000 were proven to have committed another offense within 1 year.

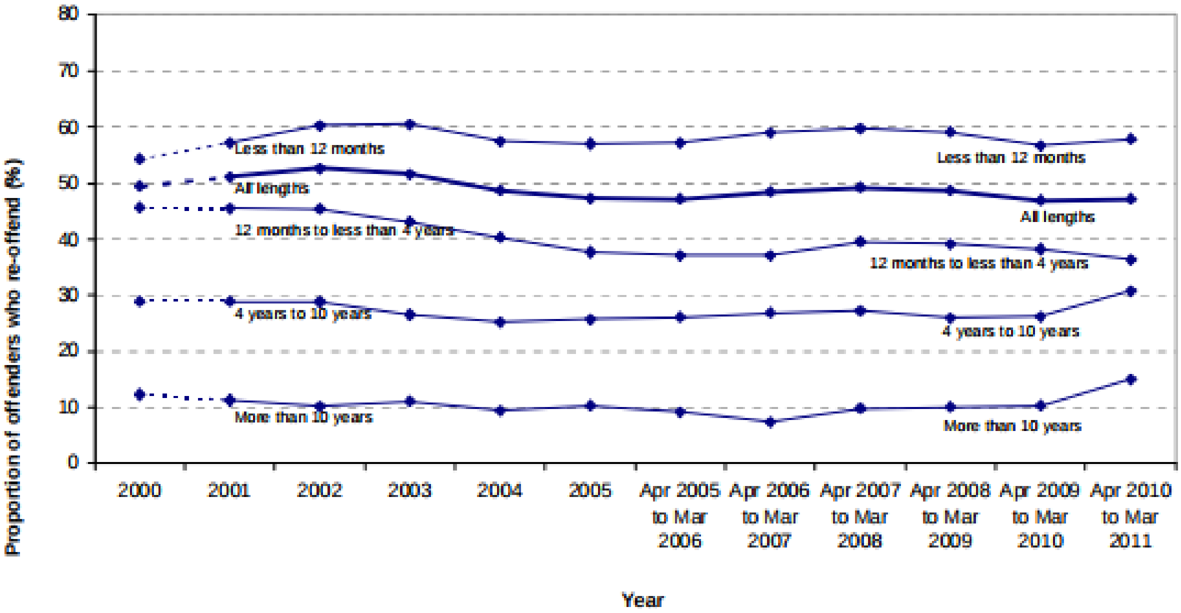


Figure 4.16: Recidivism rates for adult offenders in England and Wales

Figure 4.16 shows the population of ex-offenders by custodial sentence length proven to have committed a re-offense. From the plot, it is observed that as at the year 2000, almost 30% of re-offenders rate was recorded for the sentence length of 4 years or more, but at April 2010 to March 2011, there was a slight increase in percentage. Thus there is tendency that the re-offenders be sent back to prison again, thereby increasing the prison population.

Thus, the simple stock flow model can be modified to include the number of ex-prisoners who finds their way back into the prison, thereby increasing the prison population. I would undertake further work to implement it as future research. Hence equation 4.15.1 can then be modified by adding another variable to give;

$$P_{t+1} = IP_t + RP_t + P_{t-1} - OP_t \tag{4.16.1}$$

Where RP_t is the Recidivism population. Hence, model 1.3 can be modified to the model below.

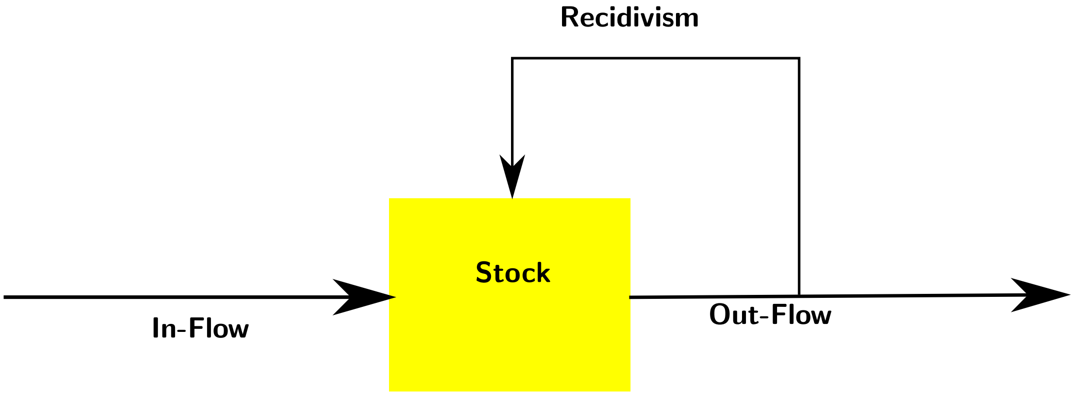


Figure 4.17: The Stock Flow Model of the Prison Population with Recidivism

Recidivism, one of the most fundamental concepts in criminal justice can be caused by many factors which include;

(a) **Unemployment:**- The absence of employment is a consistent factor in recidivism and probation violations, also, an ex-convict finds it difficult to gain employment opportunities. Research shows that 1 year after release, at least 60% of ex-prisoners were unable to gain employment nationally, according to a study by Bushway and Reuter

(b) **Love:**- Constant love and care from family is another main factor in the formation of individual and social especially for ex-convict. Prisoners' recidivism rates are associated with the amount of love, care, and support gotten from family, friend and peer groups.

(c) **Socio-Economic Factors:**- In terms of gender, men are more likely to return to prison because of criminal peer associations. e,g carrying weapons, alcohol abuse, and aggressive feelings. Also, age is another factor. According to the United States Sentencing Commission 2004, research shows that youths are more likely to offend than older people. Out of all offenders who are under age 21, the recidivism rate is 35.5%, while offenders over age 50 have a recidivism rate of 9.5% .

In addition to what has been listed above, other contributions of recidivism are poverty, lack of income, poor living conditions, divorce or separation, and death of one or both parents are other pushing factors for recidivists to re-offend and other psychological problems like anxiety, depression, addiction, aggressiveness, lack of adequate rehabilitation and reintegration services are other factors contributing for recidivism.

5. Conclusion

After the analysis performed, conclusions were drawn on the findings. Possible recommendations were provided as well as further research on the work done.

5.1 Conclusion

The aim of this study was to predict the prison population using statistical methods. Three methods were used to predict the prison population **Regression Analysis**, **Time Series Analysis** and a **Stock flow Model**. For the regression analysis, a linear model was considered, while a ARIMA model was considered for the time series analysis. For the stock flow model, we considered the simple model without recidivism after which we went on and developed a more complex stock and flow model with recidivism.

The prediction performance of regression and arima model was improved through spline and seasonal and nonseasonal differencing respectively. The results presented in table 4.14 show that the Arima Model provides the best prediction performance because its RSS is smaller, furthermore, Ljung-Box test was performed to test if the residual ACF at different lag times was significantly different from zero, In addition, the normality of the model residuals' distribution also confirmed the model's fitness. After this, a forecast for future values from January 2017 to March 2019 was made. The result were presented in table 4.14. Hence, our result reveals that ARIMA (0,2,1)(0,0,1)[12] model was appropriate for predicting the prison population of sentence length of 4 years or more(excluding the indeterminate). The stock flow model has its uniqueness as it takes into consideration how the offenders are coming in, their sentence, and how they are released, which was not as in the case of regression and arima model. These methods can be applied to a different part of the prison population but different parameters would be needed.

5.2 Future work

This section entails some possible improvements to the models and the general approach to enhance our work. For this study, there are three main suggested extensions to this work.

The first extension is the choice in linear regression methods. For this study, we used the natural cubic spline to account for seasonality and trend in the time series data, more research can be conducted to better improve the model for better predictions.

The second extension is the ARIMA model, from 4.13 it is observed that the forecast confidence interval tends to deviate from the forecast point with time, Hence the model can be improved upon for long time predictions.

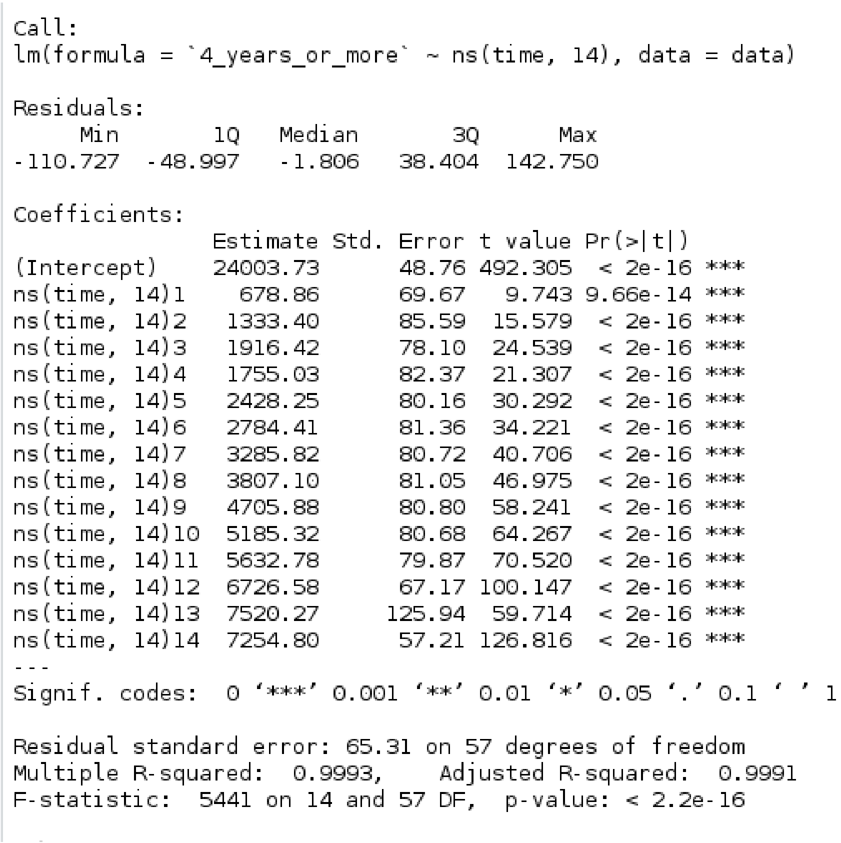
The last extension is on the stock flow model, as illustrated in figure 4.17, there is a need to introduce the percentage of recidivism into the model developed in 1.3. with an addition of the recidivist, we can better estimate the end stock. Also considering the distribution of the outflow (OP) and the Inflow (IP) and then simulation it can also tell a lot about the prison population.

5.3 Recommendation

The SARIMA (0,2,1)(0,0,1)[12] model has been designed to help policy makers and planners to help evaluate the effect of different policies on the numbers of offenders coming in and out of the prison population. Users of the tools provided above can make informed judgments about the consequences of an increase or decrease in the prisons population. The stock flow model shows how sentence lengths (which determines the stock) and recidivism are dynamic, factors outside of prison might play a vital role in the reduction of recidivism than prison itself. This is because constant follow-up efforts made on ex-prisoners might have a greater rehabilitative effect.

Furthermore, aside from the prisons population forecasting has always been very important in decision making almost at every level and sector of the economy, especially in price forecast, which is very critical to the market participant making production and marketing decisions. Early prediction about the probable price of a commodity will help the policymakers regarding the probable fluctuation in market price. this would help make proper monitoring and planning. Policy makers can get prior information about the possible future prices through price forecasting by time series analysis using the (ARIMA) model which is one of the popular forecasting model. The results from the time series analysis above can help give valuable information when formulating future policies.

Appendix A. Some additional data



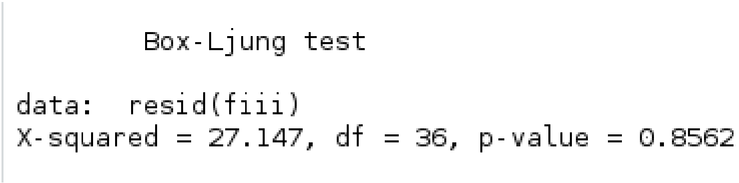
```
Call:
lm(formula = `4_years_or_more` ~ ns(time, 14), data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-110.727  -48.997   -1.806   38.404  142.750 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    24003.73      48.76 492.305  < 2e-16 ***
ns(time, 14)1    678.86      69.67   9.743 9.66e-14 ***
ns(time, 14)2   1333.40      85.59  15.579  < 2e-16 ***
ns(time, 14)3   1916.42      78.10  24.539  < 2e-16 ***
ns(time, 14)4   1755.03      82.37  21.307  < 2e-16 ***
ns(time, 14)5   2428.25      80.16  30.292  < 2e-16 ***
ns(time, 14)6   2784.41      81.36  34.221  < 2e-16 ***
ns(time, 14)7   3285.82      80.72  40.706  < 2e-16 ***
ns(time, 14)8   3807.10      81.05  46.975  < 2e-16 ***
ns(time, 14)9   4705.88      80.80  58.241  < 2e-16 ***
ns(time, 14)10  5185.32      80.68  64.267  < 2e-16 ***
ns(time, 14)11  5632.78      79.87  70.520  < 2e-16 ***
ns(time, 14)12  6726.58      67.17 100.147  < 2e-16 ***
ns(time, 14)13  7520.27     125.94  59.714  < 2e-16 ***
ns(time, 14)14  7254.80      57.21 126.816  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.31 on 57 degrees of freedom
Multiple R-squared:  0.9993,    Adjusted R-squared:  0.9991 
F-statistic:  5441 on 14 and 57 DF,  p-value: < 2.2e-16
```

Figure A.1: The summary of the Linear Model



```
	Box-Ljung test

data:  resid(fiii)
X-squared = 27.147, df = 36, p-value = 0.8562
```

Figure A.2: The Box-Ljung Test for the Time series model

Acknowledgements

My deepest gratitude goes to the Lord God Almighty for life, grace, and capability to complete this study.

I would like to express my heartfelt appreciation to my amazing, king and supportive supervisor, Sam Cuthberson, a Senior Statistician in the UK Civil Service, who despite his busy schedule was always open whenever I ran into trouble spot about my research, who consistently allowed this paper to be my own work, but steered me towards the direction whenever he thought I needed it.

To my lovely Parent and siblings Mrs. Sarah Bassey and late Mr. Samuel Bassey, Doris, Aniefiok, usoro, and Charity for their moral care and support, I say thank you. To my wonderful friends and colleagues Shadrak, Moses, Joel, Francisca, my roommate - Iris, Steven, Seth, Djibril, Brenda, Leah, Theo, James, Moshood, Sulyman, Ibrahim, the list is endless, your support was tremendous. To Samuel Afolayan my coach, you are indeed a blessing.

I would also like to acknowledge the lecturers and staff members at AIMS-Cameroon, who have intellectually imparted me during my study period in AIMS. I am gratefully indebted to my tutor Doctor Hans for his guide and valuable comments on this thesis.

References

[1] Charis Chanialidis . Statistical modelling session 4: Summarising a fitted linear model (and model assumptions). file:///home/basseyblessing/Downloads/3RD%20BLOCK/DATA%20ANALYSIS%20FOLDER/04_Friday_Model_Assumptions_slides%281%29.html#%281%29. Accessed: May 9, 2019.

[2] Chegg Study, (2019a). Statistics how to. https://www.statisticshowto.datasciencecentral.com/autoregressive-model/. Accessed: April 26, 2019.

[3] Chegg Study, (2019b). Statistics how to. https://www.statisticshowto.datasciencecentral.com/arma-model/. Accessed: April 27, 2019.

[4] (STATA 13. Arima models identification). https://www.youtube.com/watch?v=Rd67Tin8igA. Accessed: May 9, 2019.

[5] stock overflow company (2019). Stock validation. https://stats.stackexchange.com/questions/77248/what-is-autocorrelation-function. Accessed: April 28, 2019.

[6] (The pennystate university. Applied time series analysis). https://newonlinecourses.science.psu.edu/stat510/lesson/4/4.1. Accessed: May 20, 2019.

[7] Tim Bock. Display r. https://www.displayr.com/autocorrelation/. Accessed: April 28, 2019.

[8] Wikipedia. Test statistics. https://en.wikipedia.org/wiki/Root-mean-square_deviation. Accessed: May 7, 2019.

[9] Chun-pong Sing. Stock-flow model for forecasting labor supply.

[10] Department of Mathematics and Applied Mathematics, University of the Western Cape. A compartmental model to describe the population dynamics of tb disease in prisons. Accessed: April 20, 2019.

[11] Elizabeth Steiner. Estimating a stock-flow model for the swiss housing market.

[12] EVANS SKOVRON, S., E. SCOTT, J., and T. CULLEN, F. (1988). Prison crowding: Public attitudes toward strategies of population control. *Journal of Research in Crime and Delinquency - J RES CRIME DELINQ*, 25:150–169.

[13] Gov.UK (2018a). Guide to offender management statistics quarterly: January to march 2018. https://www.gov.uk/government/statistics/offender-management-statistics-quarterly-january-to-march-2018. Accessed: April 24, 2019.

[14] Gov.UK (2018b). Offender management statistics quarterly: January to march 2018. https://www.gov.uk/government/statistics/offender-management-statistics-quarterly-january-to-march-2018. Accessed: April 16, 2019.

[15] Jewel Goy (2011). Forecasting and model development. Accessed: April 20, 2019.

[16] Lowitz, R. (2013). Efficiency of life in prison in terms of human adaption.

[17] Moj (2018). Stock and flow diagrams. https://www.google.com/search?q=stock+flow+diagram+for+prisons+population&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjj8euO0tbhAhUi-YUKHfUcDSsQ_AUIDigB&biw=1869&bih=954#imgrc=eS3P6vBBSUkOzM:. Accessed: April 16, 2019.

[18] Rencontres Internationales (2002). Changing minds. http://changingminds.org/explanations/needs/prediction.htm. Accessed: April 19, 2019.

[19] Sarah Armstrong(University of Glasgow) Elizabeth Fraser (Scottish Government analytical services) (2012). Prison population projections a cautionary perspective crime and justice statistics user day march 2012. https://slideplayer.com/slide/3544014/. Accessed: April 20, 2019.

[20] Sematech (2003). Engineering statistics handbook. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm. Accessed: April 26, 2019.

[21] Soundience (2018). prisons and society. http://www.soundience.com/blog/prisons_and_society. Accessed: April 16, 2019.

[22] Transentis Consulting (2018). Stock and flow diagrams. https://www.transentis.com/step-by-step-tutorials/introduction-to-system-dynamics/stock-and-flow-diagrams/. Accessed: April 16, 2019.

[23] Wikipedia (2009). Time series. https://en.wikipedia.org/wiki/Time_series. Accessed: April 26, 2019.