UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Machine Learning Approach to Model Internal Displacement in Somalia

Sofia Kyriazi
M.Sc. Thesis
May 2018

Supervisors:
dr. Poel, Mannes
dr.ing. Englebienne, Gwenn

Human Media Interaction
Faculty of Electrical Engineering,
Mathematics & Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

At this point in time I would like to say a few words to everyone involved in this project

- It has been a great honor working for and with the United Nations;
- I hope the effort made so far can have an impact in the future of the agency of UNHCR;
- and I would also like to acknowledge the power that Machine Learning has and how it will morph our world into the future, also proud to be part of this wave.

Summary

This master thesis has been commissioned by UNHCR, the United Nations Refugee Agency, to research the possibilities of creating a predictive engine of internal population displacement within the region of Somalia and its neighboring countries. The project is directly assisted by a team of two people, the data scientist that is in charge of allocating data, detecting the sources and reporting monthly to the operations in the field, such as camps in Somalia, and Ethiopia as well and the computer science student/developer from the department of Human Media Interaction.

The first chapter will provide some background information on UNHCR, the Somalia situation and define the motivation behind predicting. At the end of this chapter we will explain the collaboration between the two main members of this project. In the next section of the thesis we will describe UNHCRs motivation for the commissioning of this thesis, the research questions created to answer the overarching research subject, the scope for this thesis. The research questions regarding the possibilities of making predictions and the best approach on exploring the data, aiming on an artificial intelligence solution that will lead to the deliverables from this research, and provide an outline for the remainder of this thesis.

Within the theoretical framework chapter, we will describe the results of the literature search. The goal of this literature research is threefold: (i) explore the pre-existing adaptation of machine learning related to population movement (ii) introduce the main machine learning techniques that will be used in the next chapters on the data collected, describing the Somalia situation (iii) techniques to perform analysis of the results and existing methodologies to select the most reliable models. The insights gained from this literature research will be used to form the methodology, which will also be described at the final section of the chapter.

In the follow-up chapter, we will explore the machine learning approaches, the methods used, Genetic and Evolutionary Algorithms and Neural Networks, and the models that were formed, explaining the process for developing each model and justifying the choices made on the selection of the training set, the testing set and the validation set. Collecting the results, we will perform comparison of the predictions.

The last chapter, will include observations on the behavior of the models and a methodology for selecting the most influential factors that affect the displacement of the People of Concern (POCs), based on expert opinions and statistical methods.

To sum up the thesis, the last part of the main body will include, observations and the discussion on the feasibility of movement prediction, where the effort should be focused and suggest possible adaptations of our methodology to operations in other emergency situations and countries.

Contents

List of acronyms

UNHCR	United Nations High Commissioner for Refugees
IOM	International Organization for Migration
POCs	Persons of Concern
AWD	Acute Watery Diarrhea
HoA	Horn of Africa
ML	Machine Learning
LSTM	Long Short-Term Memory
GEA	Genetic Evolutionary Algorithms
RNN	Recurrent Neural Networks
NN	Neural Networks
TSF	Time Series Forecasting
GRNN	Generalized Regression Neural Network
KNN	K Nearest Neighbor Regression
ARIMA	Auto-Regressive Integrated Moving Average
BIC	Bayesian Information Criterion
RVRs	Real Value Representations
ANN	Artificial Neural Networks
IDPs	Internally Displaced People
SSE	Sum Squared Error
RMSE	Root Mean Squared Error

NMSE Normalized Mean Square Error

PRMN Public Report Migration Numbers

MLPNN Multi Layer Perceptron Neural Network

MAPE Mean Absolute Percentage Error

MSE Mean Squared Error

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

Chapter 1

Introduction

International migration is a complex phenomenon, and in the recent years there has been detected an increase in migration and displacement occurring due to conflict, persecution, environmental degradation and change, and a profound lack of human security and opportunity. Migration is increasingly seen as a high-priority policy issue by many governments, politicians and the broader public throughout the world. The current global estimate is that there were around 244 million international migrants in the world in 2015 [1]. The great majority of people in the world do not migrate across borders; much larger numbers migrate within countries. There are more than 65.6 million people who are forcibly displaced around the world. Out of the 65.6 million, 40.3 million people are internally displaced within the borders of their own country and 22.5 million seek safety crossing international borders, as refugees. With the increase of violent conflict and other conditions that exacerbate forced displacement, this figure is estimated to rise in the upcoming years.

1.1 Motivation

In the Horn of Africa (HoA) [2] a situation of crisis has occurred for a time period of 7 years. In the following paragraphs, part of the motivation, we will provide a short description of this humanitarian emergency situation in the country of Somalia.

Some general information about Somalia is that the countrys total population is nearly 11 million people, and the country is divided into 18 official regions. The main source of welfare is farming, goats and camels and a funny fact is that all transactions are made via mobile phone, so there is no use of actual cash for transactions. Somalia is in the list of the most dangerous countries in the world, due to war ever since 1991, for this reason the state services are crippled. The destructive drought in 2011 has had catastrophic outcome for the farmers and their families. Almost 1 million Somalis moved internally in the country, running away from war and drought, while some of them even moved to Kenya or Ethiopia. People now live in, the temporary in the beginning, permanent now, refugee camps. Access to clean water is limited and the consumption of polluted water brings high risks for the Persons of Concern (POCs) in the camps and causes outbreaks of Acute Watery Diarrhea (AWD) and Cholera.

This short description can be expanded but it contains all the elements of the reality and the factors that affect the economy. At the same time it gives a clear image of the need for families to be constantly moving from region to region, in order to survive, leading to extreme numbers of POCs. This huge humanitarian emergency calls for humanitarian response, to save peoples lives, funding is required to prevent the situation from getting worse.

The international community reacted generously to the escalating needs in the Horn of Africa in 2017, substantially increasing funding for the responses in Somalia and the neighboring countries. Overall, more than $\$3.5$ billion was required for humanitarian action across the HoA in 2017. However, while the need of Ethiopia and Somalia got covered, Kenya was largely underfunded, resulting to limited refugee response. Another fact is that some sectors were significantly underfunded, such as Protection and Shelter in Somalia and Education and Emergency Shelter in Ethiopia[2]. With needs in the region remaining high in 2018, timely funding is required to prevent a deterioration in the humanitarian situation.

In Somalia, aid agencies were able to provide life-saving assistance and livelihood support to more than three million people per month, which helped avert famine and contain major diseases such as AWD/Cholera and Measles. A new Humanitarian Response Plan is needed, this would be an extension of last years famine prevention efforts and prioritizes immediate relief operations to help the most vulnerable, such as the internally displaced, women and children. Knowing beforehand what to expect, operations could respond in time, allocate the resources and manage the situation, to prevent diseases and improve the lives of POCs.

1.2 Framework

The need for prediction in order to assist the operations on site, triggered the research conducted and presented in this master thesis. The actual spark of the research, was the collection of interviews of POCs in Somalia. After careful examination, the data scientist of the Innovation Unit of the United Nations High Commissioner for Refugees (UNHCR) agency, remarked that the locals would sell their goats when planning to depart from a region, basically collecting economic resources to undertake the move. This observation lead to the assumption that the economic factors recorded in a region of Somalia can indicate movement of POCs.

The above observation alongside with the availability of data, describing different aspects of the situation in Somalia, including the economic components, lead to the belief that there must be a mathematical model that can describe the movement of POCs from one region to another region in Somalia or even across the borders. Reports such as [3] support that climate change also increases conflict between terrorist groups, which leads to increase in migratory flows. The millions of people facing starvation, are driven to flee also due to patterns of drought, caused by the climate change and instability. [3]-Climate change to affect migration-. The last mentioned paper, refers to extreme climate conditions to affect local migration, to pursue better living conditions, especially if the weather conditions are too extreme for the locals to adapt to them.

In our research we will focus on creating models based on the data describing the conditions in our country of concern, that can predict migration flow to a region. In a machine learning context, if the correlations between the data we provide are strong, it could lead to accurate predictions. In our case we have many potential causes for the migration effect, lots of messy complicated data, and we want to test whether a machine learning technique can lead to an accurate prediction. Several machine learning techniques will be tested, the results of which will be compared, with the final goal to select the most accurate model.

The above presented research was conducted at the UNHCR agency, with the collaboration of the Somalia Information Managers. The Information Managers located in different camps in Somalia, have had the role of collecting and reporting arrivals in each of the regions. Many other sources, described in the next chapters provide us with insights of the situation in the different states. The country of Somalia is officially divided to a number of 18 states, and POCs flee from state to state, as well as the camps located near the borders of Ethiopia. To narrow down the scope of the thesis, the research will target making predictions, using machine learning, in one of the regions of Somalia. These predictions will portray numbers of arrivals in a region for the upcoming month, based on the data collected, reflecting the situation as reported by the Information Managers, and the rest of the data sources, which will be introduced in the next chapters.

1.3 Research questions

In this section, the research questions, addressed in this thesis will be established, aiming to provide a structure to assist the overall framework described in the previous sections. These research questions will attack the main research question of this thesis: *How can machine learning assist in predicting human displacement in the country of Somalia?*

To structure this research, and consequently the next chapters of the thesis, we present the following research questions:

1. Between Regression Machine Learning approaches and Genetic and Evolutionary Algorithms, which ones can provide predictions for arrivals of POCs in the Bay state of Somalia?
    (a) How do Genetic and Evolutionary Algorithms perform?
    (b) How do Recurrent Neural Networks perform?
    (c) What are the measures of the performance of our models and how do we compare these models?
2. Which are the most influential variables in the Models that were developed?
    (a) Most influential variables for Genetic and Evolutionary Models
3. What are our observations and conclusions from the results of our experiments?

1.4 Report organization

The remainder of this report is organized as follows. In Chapter 2, we shall provide the literature review and describe the theoretical framework, as a basis for our experiment. The literature review will highlight basic information of efforts made so far in the field of predicting migration and the machine learning techniques we will apply later on, and annotate the most important and mutual elements in preceding research that examines movement flows. And ultimately, position this research in the broader field it belongs in.

Then, in Chapter 3 we will present the Genetic and Evolutionary Algorithms and Neural Networks adaptations to our data, the limitations and challenges, examine both models, and keep as a baseline Linear Regression. In that chapter we will define the methodology and justify our experimental set-up. Later in the chapter of Results we will measure the performance of the methods used, compare the results and conclude on the best approach of Machine Learning (ML). In the next Chapter 5 we will discuss a possible combination of the models, try to detect the most influential variables for the outcome of the models, and cross validate the assumptions with the experts' opinions.

Finally, in last Chapter 5, we shall discuss all the conclusions from the model testing and give recommendations for adaptation of our methodology, for expanding predictions in the rest of the states of Somalia or even in neighboring countries.

Chapter 2

Theoretical Framework

2.1 Predicting Migration

Predicting migration intents to project flows of population, and usually migration is connected to relocation across the border of the country according to the Migration Data Portal [4]. This concept is limited, as mentioned by the official researchers, in the International Organization for Migration (IOM) report [5], by some of the following factors, which are also going to influence our research and methodology:

1. The definition of migration varies according to the country, due to the dissimilarity of the motives, driving factors for displacement.
2. Access to data, and availability, is often hard to capture and restricted.
3. Accuracy and detail in data, are considered a luxury, due to how uneventful the need to collect data is before the actual displacement occurs.
4. The many theories developed to explain migration have failed, due to the inability to interconnect the *push and pull* factors.

For the modeling of international migration, there have been many efforts to describe the phenomenon, combining different disciplines from demography, economics and sociology. The research by Bijak [6], describe the theories developed so far, as well as some theories that unify these *disassociated* theories. The microeconomic theories, that treat migration as a result of the cognitive process of a cost-benefit analysis of the individual, offer an optimistic approach, in which the economic factors of giving - receiving migrants countries can be measured, and the decision is based on a maximizing-minimizing function.

In the same paper [6] it is argued that internal displacement, implicates different criteria for the decision to flee, since it lacks in institutional restrictions, the geographic theories can better interpret the criteria for migration. Geographical theories, account the distance between the region of origin and destination, or the cost of transportation, and weights for that model are considered the economic factors, neglecting that dynamic systems as such, may be undergoing qualitative changes.

In the paper by Kupiszewski [7], claims that theories can be used post-migration, to explain the phenomenon but they cannot be used to forecast the population flow, since some of the theories are to complex to be expressed in mathematical terms or that simple theories cannot accurately describe it. Nowok as well as Kupiszewski [8] support that macro-level statistics are usually incomplete and have deficiencies, so they cannot serve for predictions on a large scale. In our case, we argue that we try to collect macro-level statistics for each region and predict for a smaller scale, and use machine learning to approach mathematically-based approaches without depending on the experts to create a theory that interprets the Somalia migration phenomenon.

In the following limited section we will examine related work to the framework of this thesis, not for migration, but for unpredicted flow of population, either that is related to tourism (driven by different factors), or even hospital emergency overcrowding.

Economic factors and methodology of approach in the following paper [9] signifies that in tourism many methods have been used to make predictions. The training data set includes economic factors from the place of origin to the place of destination, as well as hotel prices, but the difference is that the place of destination performs some campaign to inform tourist and promote the destination country, while we have to assume that in our case the migrating population seeks for that information from alternative sources. All the algorithms tested seem to be working quite well for small amount of data on Tourism Forecasting, but the MLP had a relatively bad performance compared to the alternative algorithms that were tested such as Generalized Regression Neural Network (GRNN) , and the K Nearest Neighbor Regression (KNN). In the same paper, the notion of time t is introduced as part of the dataset for the training. In general the GRNN seems to be performing better for all the datasets tested, and the interesting point the paper makes, is the unexpected Asian crisis that affected the tourism, which complicates making predictions. Forecasting as well falls under three categories as accurate, good and inaccurate forecasting according to the Mean Absolute Percentage Error (MAPE) values.

Another case that we can use as an example, to guide our research ,is the modeling and forecasting of arrivals in a hospital's emergency room, as described by Kadri [10]. The motivations of prediction comply with ours, forecast demand in emergency departments has considerable implications for hospitals to improve resource allocation and strategic planning. An autoregressive integrated moving average Auto-Regressive Integrated Moving Average (ARIMA) method was applied separately to each of the two categories of data of total patient attendances, as described in the paper [10], that lead to optimistic results, even thought the simplicity of the experiment. The data window needed to train the model was relatively small, with $t = 168$ the model had a good fit.

The above refer either to emergency situations, or are similar to migration because they concern population flow from point a to point b. In migration prediction, in the paper by Simini [11], the radiation model, is based on population distribution and the distance between the point of origin to the point of destination. Another source [12] by Lenormand on trip laws, makes use of the gravity model, and concludes to distance having more impact on movements, such as migration, than the opportunities in the point of destination.

Finally, there has been related work, in the paper by Robinson and Dilkina [13] that was published on November 2017, after the initiation of this project, which also suggest measures for comparison of the predictions made with machine learning techniques. This paper, suggest that in order to make predictions for more complicated dynamics we can make use of machine learning models. The paper focuses on making predictions, of migration, between the states of the US.

Some of the measures mentioned in this paper, and used for comparison, are the Mean Absolute error (Mean Absolute Error (MAE)), the goodness of fit (r^2), the root mean squared error (Root Mean Squared Error (RMSE)), and a similarity score, to compare the results of the predictions with the actual arrivals. Some of these measures we will also be using later on 3. In the research mentioned above, the XGBoost Model is used as well as an Artificial Neural Networks (ANN) Model, the results of which are compared at the end. The XGBoost Model, is preferred as it allows to detect feature importance. Identical to the research that we will perform and we will base our methodology on this paper.

2.2 Multivariate Time Series

Forecasting is the prerequisite for making scientific decisions, it is based on the past information of the research on the phenomenon, and combined with some of the factors affecting this phenomenon, proceeding by using scientific methods to forecast the development trend of the future, or in simple words it is an important way for people to know the world. [14]

To define forecasting, in the scope of our project, we aim to make forecasts to influence decisions on planning and preparedness. Our methodology will be based on past observations of the phenomenon of migration in Somalia, and combined with the factors that possibly are affecting this phenomenon. The scientific methods we will use for forecasting, will be described further in this research report.

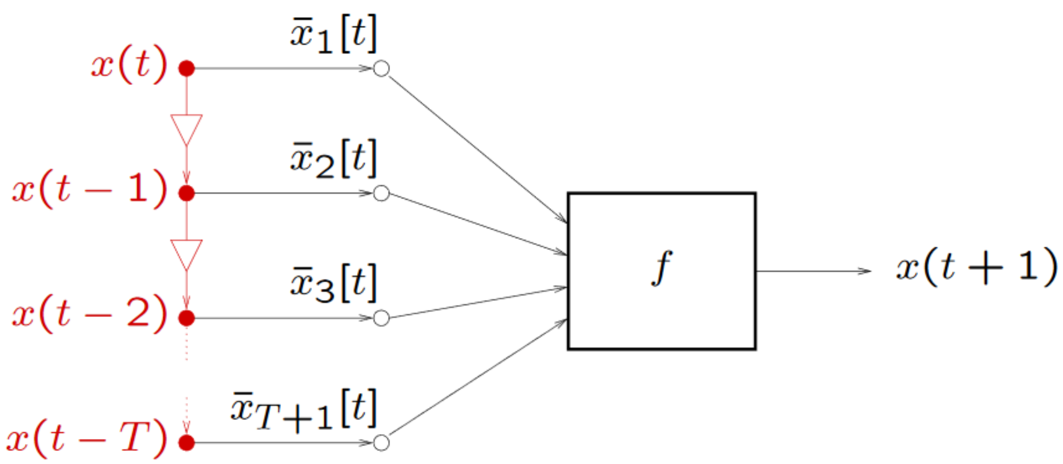

Figure 2.1: Time Steps Lag for Multivariate Time Series Forecasting

Defining forecasting: A time series is a sequence of vectors $x(t)$, where $t = 0, 1, ..., T$ and t represents time. We consider now to be represented by t=0 and the next time point t_+1. Vector x is a multi data point vector, and contains all observations of influencers for our phenomenon at time t.

To introduce the term Multivariate we will describe the vector x. Let us consider n observations recorded at time t, then for each t we derive an vector of $\{x_(t1), x_(t2)...x_(tn)\}$. resulting in a matrix such as we can see in the matrix below.

$$\begin{bmatrix} x_{01} & x_{02} & x_{03} & \dots & x_{0n} \\ x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ & & \dots\dots\dots & & \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

In the [15] time series analysis is defined to belong in the following categories:

1. Forecasting of the future development of the time series.

2. Classification of time series, or a part, into several classes.

3. Description of a time series in terms of the parameters of a model

4. Mapping of one time series onto another.

For our research we will focus on categories $1\&3$. But let us first give definitions for two important notions, commonly used in time series forecasting. The term *lag* and the term *sliding time window*.

The lag operator d otherwise known as back shift operator d, is the shift of a time series such that the lagged values are aligned with the actual time series. The lags can be shifted any number of units, and the units are determined by the time series themselves, in our case the unit is a month. We can restructure any time series dataset as a supervised learning problem by using the value at the previous time step to predict the value at the next time-step. The use of prior time steps to predict the next time step is called the sliding window method. For short, it may be called the window method in some literature. In statistics and time series analysis, this is called a lag or lag method.

Category 1: Finding a function R, such as to obtain an estimate at time t+d, given the values of x up to time t-l, where l, is going to be under investigation. Variable d is often defined as the lag of prediction, in our case we will focus on d=1, but it could potentially be that d obtains a higher value. In practice that would mean, if d=2 then perform prediction two months ahead, of any displacements to a region in Somalia.

Category 3: Looking closer to the function R, we aim to detect the most important influencers, affecting the movement, therefore human decision of IDPs in one region of Somalia. This will help eliminate parameters, and we expect the function to have fewer parameters than the input vectors,to help us understand and describe our time series.

The most commonly used methods for Time Series Forecasting as mentioned in [16], are the Exponential Smoothing and the ARIMA model. These techniques require some preprocessing on the data to detect seasonality and trends. The metrics used in Time Series Forecasting (TSF) to measure the accuracy of a model, are the Sum Squared Error SSE, the Root Mean Squared and the Normalized Mean Square Error, as denoted in the [17]. We will examine evaluation methods, in the chapters to follow.

2.3 Machine Learning

2.3.1 Genetic and Evolutionary Algorithms for Time Series Forecasting

The genetic and evolutionary algorithms GEA, are considered a novel technique for Machine Learning tasks and an alternative to simple regression. Linear regression is an attractive model, commonly used in machine learning because the representation is, so simple, a linear equation that combines a specific set of input values $x_1, x_2...x_n$ the solution to which is the predicted output y. Let us give a short description of the GEA and then focus on a category of this class of algorithms. GEA are part of the Evolutionary Algorithms, mechanisms that mimic Darwin's process of natural selection, where only the stronger input variables affect the output. The stronger input variables are decided based on the fitness of the solution to the data.

GEAs are frequently used for modeling binary input data, but since the binary encoding is usually inappropriate for real application, there is an alternative method, the GEA with Real Value Representations. This special category performs a stochastic selection, which favors some of the parameters, represented by real values, and generates a solution, that maximizes the fitness of the model. This combination is useful for numerical optimization process. In our case the numerical optimization process is the selection of some of the parameters, which are the driving factors, leading people to flee in the region of Banadir. Our parameters are not binary, they are consisted of different units, but their nature is numeric. We are trying to optimize the fitness in predictions of arrivals. In the following section we will mention the few cases where GEA with Real Value Representations (RVRs) have been applied for TSF.

GEAs are based on the natural selection process of selecting the optimal solution, given the source variables and the target. The first step of the GEAs is random sampling, which means that different runs of the same program will produce, alternative solutions, with different influencing variables. They are considered non-deterministic approaches, it exhibits different behavior, in contrast with all the regression algorithms, as we described before, that yield the same result, if the setting of the run remains intact.

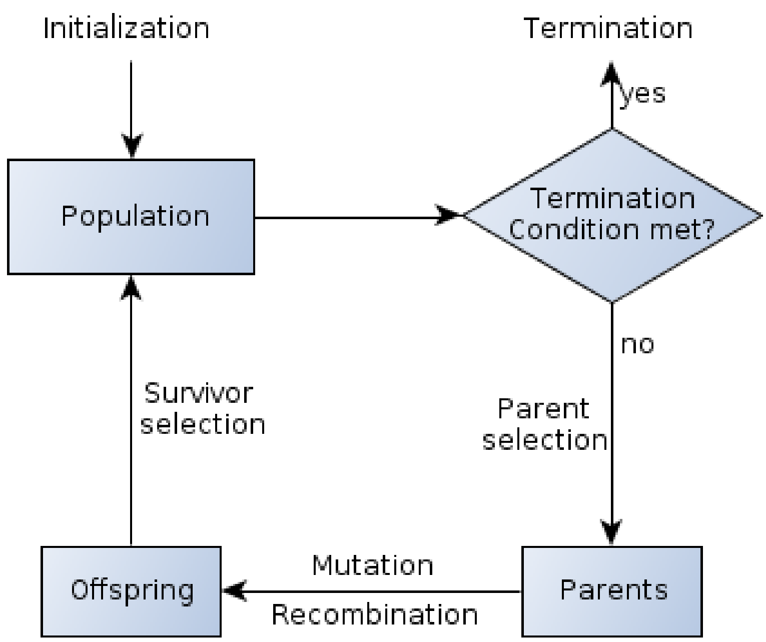

Figure 2.2: The flowchart of the processes in GEAs

Genetic and Evolutionary Algorithms for time series forecasting, has been applied to predict air pollution [18]. The GEA are used to design an architecture for predicting concentrations of nitrogen dioxide at a traffic station in Helsinki. The GEA is compared to a Multi Layer Perceptron Neural Network (MLPNN), and they doth try to deviate from the common techniques such as regression, to approach this much more complex phenomenon. The justification for using GEAs is the high dimensional sample space, and the *chaotic* relations between the input variables. Another method used to decrease complexity, on a technical aspect, is the use of parallel processing, this will play an important role on our Discussion chapter.

Another common use of GEAs is financial forecasting, [19], where GA are used to obtain time-series forecasting rules for macro-economic figures. The GEAs show promise, as they over perform more traditional methods, in forecasting, such as the ARIMA. For model selection, the Bayesian Information Criterion (BIC) criterium is used, the same criterium we will include in our methodology as well.

2.3.2 Neural Networks for Time Series Forecasting

The main advantages of Neural Networks are that they have the ability to learn and model non-linear and complex relationships, which applies to our case where the inputs and the target are non-linear and complex. Also Neural Networks (NN) have the ability to generalize, so we can expect our model to be able to predict on unseen data, having the knowledge of the input training data. Contradicting other models, used in statistics, NN do not impose any restrictions on the input variables. Additionally, many studies have shown that NN can better model hidden relationships in the data without imposing any fixed relationships in the data.

Neural Networks are algorithms, intended to simulate the neuronal structure of mammals, on a smaller scale and with less processing units. Neural Networks work in layers, and use a *learning rule*, that they readjust when a *guess* their performed proved to be wrong or right. The reasoning in selecting Neural Networks for modeling migration in our case is that our problem falls in three out of the four categories that Neural Networks are most commonly used as we can see below:

1. capturing associations or discovering regularities within a set of patterns
2. where the volume, number of variables or diversity of the data is very great
3. the relationships between variables are vaguely understood
4. the relationships are difficult to describe adequately with conventional approaches

We cannot argue that our dataset is great in volume, as we decided to aggregate the data as we demonstrate in the next subsection, but we can claim that the number of variables is large and that the data is diverse, even though it is only numeric values, they represent different units, such as cash, river levels, number of people, incidents.

Neural Networks in their simplest format, do not take into account the time dimension, which has to be supplied in an appropriate manner. Recurrent Neural Networks Recurrent Neural Networks (RNN) are suggested to resolve this problem, as they preserve order of the input variables. Another important problem is the need to capture short or long term dependencies in the sequence of the data. A special category of Long Short-Term Memory (LSTM) Recurrent Neural Networks, can be used to capture the most important past behaviors and account the importance of these behaviors for future predictions. [20]

There are several applications where LSTMs are highly used. Applications like speech recognition, music composition, handwriting recognition, and even in as we saw in the above section there has been some use of NN for current research of human mobility and travel predictions. Recurrent Neural Networks make use of the output the model had from the input, that output is fed back as input, to generate a new output and so on, this type of NN deals with sequence problems. Recurrent Neural Networks are frequently used for speech or video processing, music composition because it is important to store knowledge of past instances, to interpret new instances of the data.

LSMT, introducing memory and the temporal dimension to Recurrent Neural Networks. This specific type of NN has been used to answer questions in clinical medical data recognition [21] that have similar characteristics to our data, in the manner that the sampling is irregular, data could be missing, and they also are interested in capturing long range dependencies.

An unexpected research that was triggered to preserve species vegetation, is the Predictions of Elephant Migration [22] also based on Recurrent Neural Networks, trying to predict a single elephant's position point. The elephants are also restricted in a reserve, and so migration is limited within those borders. This research will also guide ours, even though we are predicting massive arrivals and not individual's movements.

2.4 Evaluation Metrics

Error measurement methods define the forecasting accuracy and enact a critical role, allow monitoring for outliers in predictions, and will standardize our forecasting process. Depending on the type and volume of the data, as well as the nature of the predictions, different Error Metrics can be used in order to interpret the models derived from the machine learning process. In this section we will examine the most common error metrics, we detected in our research on the evaluation process of the papers we examined during our research. Furthermore, we shall examine the advantageous and unhelpful aspects of each metric. In the Methodology chapter we will discuss further the most suitable metrics, for our approach on arrivals forecasting.

Statisticians define evaluation as the systematic approach on a set of predictions compared against the labeled actual values and compared to come up with metrics that determine the performance of a machine learning model. The approach does not differ by a lot, on machine learning evaluation. The first step before the generation of the model, is the division of the dataset to a training set and a validation set. Machine learning models are trained on the training set and once the training is completed, the model can be used to make predictions. The validation set is used to test the already trained model with a larger subset of the original dataset. A few common metrics, used for evaluation are the following:

1. **Classification Accuracy** is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

   $$S_n = \frac{NumberofCorrectPredictions}{TotalNumberofPredictionMade}$$

   It works well only if there are equal number of samples belonging to each class, that wouldn't be therefore appropriate for our dataset.

2. **MAE** is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they dont gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as:

   $$MAE = \frac{1}{N}\sum_{j=1}^{N} |y_i - p_i|, p_i = prediction$$

   Which is more appropriate since we are requesting an approach to the number of arrivals.

3. **Mean Squared Error (MSE)** is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the square of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we

take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_i - p_i)^2, p_i = prediction$$

4. **Mean Absolute Percentage Error (MAPE)** measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error, as shown in the example below:

$$MAPE = \frac{1}{N}\sum_{j=1}^{N}\frac{|y_i - p_i|}{y_i}, p_i = prediction$$

   The MAPE is scale sensitive and should not be used when working with low-volume data. Notice that because "Actual" is in the denominator of the equation, the MAPE is undefined when Actual demand is zero. Furthermore, when the Actual value is not zero, but quite small, the MAPE will often take on extreme values. This scale sensitivity renders the MAPE close to worthless as an error measure for low-volume data.

5. **RMSE** is a standard regression measure that will *punish* larger errors more than small errors. This score ranges from 0 in a perfect match, to arbitrarily large values as the predictions become worse:

$$RMSE = \frac{1}{N}\sum_{j=1}^{N}\frac{|y_i - p_i|}{y_i}, p_i = prediction$$

6. **Sum Squared Error (SSE)** If the error is defined as the

$$e = |y_i - p_i| p_i = prediction$$

   then the SSE is defined as the

$$SSE = \sum_{j=1}^{N} e_j^2$$

   and the

7. **Normalized Mean Square Error (NMSE)** as the

$$NMSE = \frac{SSE}{\sum_{j=1}^{N}(y_i - p_{mean})^2}$$

8. **Bayesian Information Criterion (BIC)** is the penalty oriented function as we can see in the formula below, and commonly used in statistics

$$BIC = N * ln(\frac{SSE}{N}) + v * ln(N), v = variablesnumber$$

In the tourism forecasting paper [9], the metric used to evaluate the predictions for the test set, is the MAPE and the averaged standard deviation, while experiments are executed where the dataset is divided into random training sets and testing sets. While, in the more generic paper [23], regarding GEAs and time series forecasting the methods suggested are the SSE,RMSE, NMSE as well as the more statistical approach that uses BIC.

2.5 Data

This master thesis will focus on the prediction of displacement of POCs in the most dense populated region of Somalia, the region of Banadir. As we can see in the figure below, there are 18 regions in Somalia, as the country is officially divided to that number of states. These predictions will reflect numbers of arrivals per upcoming month, based on the data collected, reflecting the situation in Somalia, more on the data collection is included in the following paragraphs and the next Chapter 3.

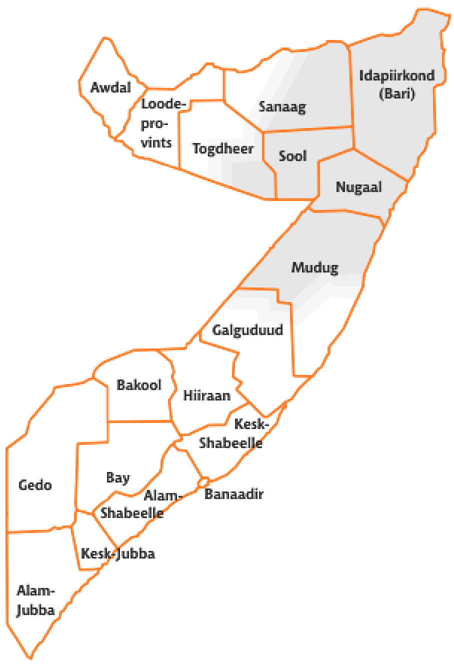

Figure 2.3: Somalia division to 18 states

We argue, that we must set our target to predict arrivals at a certain state of Somalia, as in the most valuable prediction for the information officers at the state, to improve planning accordingly. The argumentation for this geographical scope is two-folded. Each operation is located in a different region of Somalia, therefore they act individually, and ideally, to organize and allocate resources more efficiently they would like to know in advance how many POCs to expect. The original datasets, formed by the reports of arrivals in a region hold that kind of information, leading to the transformed dataset including numbers of arrivals at a certain region, for each month of recordings.

Even though the original collection of data was cleaned and transformed to daily numbers of different measurements, that from then on we decided to keep track by collecting and parsing, we decided to build the models on a monthly basis. The main arguments for the scope of predictions to be on a monthly basis are:

1. As the collection at is final format resulted in a large dataset, the software, that we have been using, needed more time to produce a model with good fit. Therefore, in order to decrease the training time and to increase the process of testing and validating the models we decided to aggregate the data, after parsing and cleaning, to a monthly scope. The impact of this decision would also be that our iteration circles would become faster, this way we could decrease any faulty assumptions or selections of models. We will elaborate more on this, in the next chapters. Basically, the re-adaptation was less time consuming if the volume of the training set was decreased.

2. If we were to predict on a daily basis, the chances to make errors would increase. Given that the model would be dealing with relatively small numbers of arrivals each day, the model fit, after the training, would approach the small numbers but the error would be relatively big.

3. Some of the data data could not be adapted to daily, such as the prices of goats and water. These numbers are collected from the official websites, and they reflect for each month what the price was for each region of Somalia. The assumption that the price was the same for every day of the month would be wrong in our case.

Our data, from enumerate sources, provide use with state based features which we will aim to combine, into hopefully interpretable models either with the use of RNN or GEAs. Our observed variables for the multivariate vectors, are chronologically ordered, and the assumption is that one of the patters that occurred in the past, will reoccur in the future. Our discrete time steps for the 7 years of data times the months for each year : $12 * 7 = 84$ vectors. The size of the time interval $= 30$ days.

Some of the data points are averaged over our time interval, for us to obtain the series. Sampling depending on the data source we have been using. The sampling frequency of these sources was daily. Each vector is described by multiple variables, so we can therefore use the term multivariate time series dataset.

As was mentioned in the paragraphs above, we decreased the volume, by aggregating to monthly data. The variety, is the next dimension, we are going to analyze. The organization of UNHCR collaborates with external organizations that collect unstructured data from internal sources such as sensor monitors, interviews, statistics from internal reports and some of the data comes from spreadsheets. The velocity of our data, the rate of which it is being produced, is the last dimension we will define. Since there are sources, per example, the click rate of a buyer on an e-shop, with high velocity, meaning data is being produced every minute of the day, we can say that we have a low frequency production of information. Different sources have different rates, for example rain is collected daily, but prices monthly.

The Extraction, Cleaning and Annotation phase [24] is the biggest part of the preparation of our data. This phase is applied with scripts, different strategies to extract information from graphs, reports in pdf format, spreadsheets and websites. Each data source is given in a different format and yet there have been cases where the format changed during the period of 7 years. For those cases we have different scripts, even for the same data source. The nature of our data is historic, so the forms of Machine Learning we will be using need to include the temporal dimension of the data.

Chapter 3

Method

All the papers that are part of the research chapter, will guide the methodology we developed to design our models and to evaluate our results, in order to answer the main research questions, as mentioned in 1.3. In the following section of this chapter we will analyze the approach we took on the data, given the availability, as well as the methodology we will follow on the machine learning aspect. In this chapter we shall address the different steps taken to perform the creation of a machine learning predicting system with focus on one state of the country of Somalia. The following cases are answered through the course of this chapter:

- Material and Data Preparation
- Machine Learning Framework
- Evaluation Framework

3.1 Material and Data Preparation

For the scope of this master thesis, we will target the Banadir region of Somalia, which also contains the capital of the country, that has the most significant number of arrivals, as is noticeable in the historically collected data. Outlying values for Banadir Arrivals in Region : peaks at $45.938, 37.053, 81.695$ and 115.474. The graph below3.1 shows, in yellow line Banadir arrivals that are the highest amongst all the regions and they peek on extreme values for some of the months, while Banadir has an average of 14.835 arrivals per month, the rest of the regions have an average of less than 4.000 and the median of Banadir is around 8.000 arrivals, making it a region of interest for predictions, according to the policy and information officers in the country of Somalia.

The calculations of the arrivals, derives from the original Public Report Migration Numbers (PRMN) files, collected by the Information officers in the field, that report in

STATE	MEAN	MEDIAN	MAX
Awdal	409	185	5,488
Bakool	1,245	253	11,200
Banadir	14,835	8,007	115,474
Bari	754	277	6,895
Bay	5,281	870	71,880
Galgaduud	4,208	1,130	36,014
Gedo	2,264	1,225	16,464
Hiiraan	3,258	266	54,400
Jubbada_Dhexe	930	582	4,294
Jubbada_Hoose	2,293	1,614	14,156
Mudug	2,153	510	61,683
Nugaal	365	194	2,759
Sanaag	2,130	213	31,996
Shabeellaha_Dhexe	2,210	660	35,430
Shabeellaha_Hoose	5,029	1,710	76,765
Sool	1,469	267	48,938
Togdheer	1,150	100	14,811
Woqooyi_Galbeed	876	316	19,698

Table 3.1: Statistics for the states of Somalia

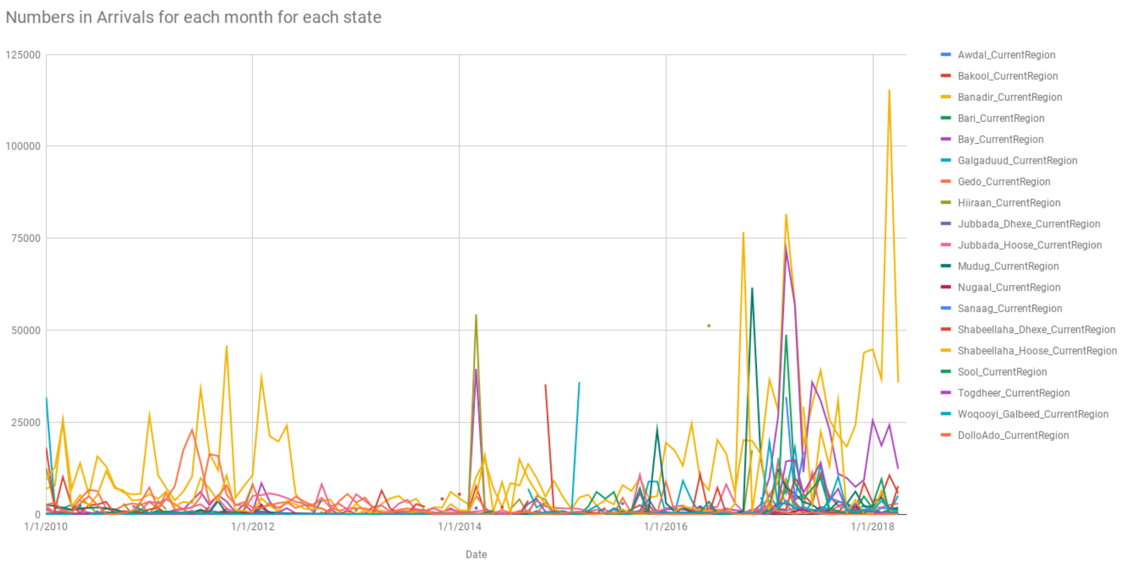

Figure 3.1: Arrivals per month in each state of Somalia

the same document, the origin of a POCs or a group of POCs, the date it entered the current settlement, and the number of people in the group as well as the reason of resettlement. Statistically we can mention that most of the cases concern economic factors of fleeing such as ”Could not afford to stay in the previous location (if IDP) or country (if cross border)” or factors related to conflict, eviction and safety.

To obtain a general view of the situation in each state of Somalia, the following data sources were collected monthly that represent some economic factors, some factors that reflect climate conditions and some socio-political factors.

The sources of data A.1 are the following:

1. PRMN dataset and we divide this information to

    - CurrentRegion, as in Number of People currently in that State
    - FutureRegion, reported Number of People fleeing to that State
    - BeforeRegion, reported Number of People fleeing from that State

2. ACLED dataset and we filter this dataset to

    - Fatalities, as the sum of the number of deaths in violent incidents, as categorized according to some internal criteria
    - Violent Incidents, as the number of incidents of violent nature such as protests, terrorist attacks etc.

3. Prices, and we collect prices that reflect the market for

- Water Drum Price, as recorded by the official website SOALIM.
- Goat Price, as recorded from the same website as a source.

4. Climate indicators, and we collect prices that reflect the market for

   - Rainfall per State, as collected by sensors in the different stations of a state, and averaged.
   - River Levels per River, as collected by sensors and averaged, by stations.

3.1.1 Limitations

Originally, we tried to collect many more data sources such as cases of AWD or deaths per region, indicators of low quality of life conditions, or even funding that the states were receiving, indicating potential for development in the region. However, these sources were not updating every month, and what we collected was scattered, hence we decided to exclude these data collections from the final collection. The collections of the final dataset we will use for the Machine Learning part of this project can be found in the A.

Another limitation is the experimentation with different states of Somalia, due to the time frame of this research. Therefore, we aimed to examine closely only one state, the state of Banadir, as it resembled much interest and it was associated with movement of POCs either in terms of arrivals or/and departures. Other areas of Somalia, with similar characteristics, would also be interesting as well to investigate further.

3.1.2 Data Outliers and Missing Values

Performing a visual overview of our data, we can detect that there are cases of outliers as well as missing values, in almost all the different data categories. Therefore, we have to justify the action to take in each case. In this subsection, we shall explain shortly how we will handle outliers and missing values. An outlier is an observation that lies an abnormal distance from the rest of the values in the dataset. [25] In our case we have, detected outliers in almost all the collections. These abnormal observations could fall under one of the following categories:

1. Outliers are the result of measurement or recording errors

2. Outliers are the unpremeditated and exact outcome resulting from the recordings

In our case outliers fall under the same category, as when there are recording errors, there are no results collected, as we will explain in the section below, of handling of missing values. Outliers could contain valuable information, especially when they regard the target, as in the number of arrivals for a state. So its important to treat our outliers as they are recorded, and we assume that these values are correctly reported by the officials. The case of Water Drum Prices being extremely high, is correlated to the conditions within a state and these are not errors, that we should ignore, but paradoxical indicators for movement, either pushing or pulling people from state to state.

In the case of missing data, we can assume that the missing value, falls under one of the following categories:

1. The sensors, which provide data for two of our categories, rain and rivers, have failed, to give feedback and they have not been replaced within the time period of the month, sufficiently for us to be able to average the values.

2. The information officers were unable to register arrivals and departures, therefore, we have missing value for the Region in terms of numbers for Current, Before and Future.

Missing values, can be treated with one of the following techniques, and we have selected to experiment with two of these, because they serve better the purpose of this project:

1. Replace missing values with 0. This would not fit the needs for our case. If we choose to replace missing values of arrivals, for example in a region, with zeros, then we are alternating the training set to model arrivals for that regions to zero, whereas other influential variables might be pointing out that there were a lot of arrivals for that region, but unfortunately they have not been recorded. We would not pursue to bias the machine in such a non-rational manner.

2. Replace the missing value with an alternative value, either that being the mean value, the median value, or the previous instance value. Again since there is not a pattern of arrivals or rainfall in the datasets with the missing values we cannot make the assumption, that we can guess the missing value. Using the rainfall in Gedo as can be seen in the table below 3.2, we cannot assume that for September there was no rainfall, or that there was the mean of these series rainfall, as the series has extreme deviations from the mean and the median, as well as the next rows don't follow the last rows value.

3. The last option is to exclude the entire row, for all categories from the training set thus leading to gaps between dates in the training set, and since in both of

Date	4/1/2017	5/1/2017	6/1/2017	7/1/2017	8/1/2017	9/1/2017
Rain	83.2	44.7	0.0	0.0	0.0	
Date	10/1/2017	11/1/2017	12/1/2017	1/1/2018	2/1/2018	3/1/2018
Rain	65.0	110.0	0.0	0.0	0.0	0.6

Table 3.2: Rainfall of 1 year of data in the Gedo state

our machine learning models, we would want to base prediction on patterns, this will make the training set smaller but a lot more reliable.

The reason we examine the training set for the next step of our methodology is that there has been research indicating a relationship between accuracy of models and outliers, as well as missing data. The paper by [26] tests on multiple datasets ANN with different percentages of missing data and concludes that potentially significant information loss is produced even with small percentages of missing samples.

For outliers in the training data, it has been demonstrated that modeling accuracy decreases as the outlying points increase. [27]. In the same paper it is concluded that when the outliers, are less than 15% of the total data then the models accuracy is statistically significant compared to having no outliers data. This study also shows that variations in the percentage of outliers and magnitude of outliers in the test data may affect modeling accuracy.

Given these conclusions, of previous researchers, we will also experiment, and compare the accuracy of our models, using the technique of disregarding outliers and including them. More on the training set will be explained in the experimental set up and the Results section.

3.2 Machine Learning Framework

This section deals with the choice of the machine learning approach. Building the Machine Learning models for our problem, we need to define the objective of prediction, and use the definition we provided for our dataset in the section above. Questions such as:

- Which algorithms are considered suitable for our dataset?
- What are the main challenges, after selecting the ML approach?
- How can we evaluate the models produced and compare the results?

3.2.1 Specifications on Machine Learning

Choosing the right machine learning approach comes with advantages and disadvantages of each technique. In spite of the many different machine learning approaches we came across during our research, on multivariate time series regression problems, we decided to focus only on the two described below, but let us first define the goal behind the model building.

Representing n zones of interest, states of Somalia, where $n = 18$, each with an array of d_{12} variables, for $t - 86$ to t time steps, the target of predictions is the d_1 variable, representing the arrivals in region n_1 at time step $t + 1$. Our goal is to generate models, that use the variables for all the zones of Somalia, and outputs the predicted number of arrivals of Internally Displaced People (IDPs) in region n_1. For our experimental set-up we will focus on the region of Banadir. Our approach works under the assumption, that arrivals in that region can be entirely based on the variables we are feeding to the algorithm from the previous time steps, and therefore predict the next month's arrivals.

To predict d_1 for $t + 1$, we will make use of the GEAs [23] and the RNNs [20].

1. **Neural Networks** Commonly used for predictions in financial time series forecasting where data volatility is very high. Given that regression forecasting problems are complex, with a lot of underlying factors in our case, where we could potential have included in the dataset the correct factors or not. Traditional forecasting models pose limitations in terms of taking into account these complex, non-linear relationships, while NN, applied in the right way, can provide us with undiscovered relationships between our data.

   Specifically Recurrent Neural Networks, consider the sequence in the training data, and can therefore respect our input, and create relationships that propagate on delay.

2. **Genetic and Evolutionary Algorithms** are applicable in problems without pre-existing methods available. Another reasoning, to support our decision, is that GEAs can deal with discontinuities and noise in data, as we mentioned in the previous section regarding our data. Also, GEAs can deal well with discrete variable space, such as ours and can incorporate *if then else* constructs. Since, GEAs produce multiple solutions, they are principally used for multi-objective problems.

For the two approaches mentioned in the above section, specific implementations have been chosen and used to predict the arrivals based on our given data set. An important aspect of using these approaches is the tuning of different parameters. Regarding the GEA, an evolutionary evaluation is performed, by building certain amount of instances of the algorithm, with a dataset including data up to two distinct historical points. A fixed number of stability of the model is used to determine the finalization of a solution, after all the training set is covered and the fitness of the model remains stable. Models are produced with the same set of component functions, which we will describe in the experimental setup section before we present the results. The fitness measure used, is the MSE, and the training set is split to 90% training set, by randomly selection.

The software used for the implementation of the GEA is the A.I. powered modeling engine created in Cornell's Artificial Intelligence Laboratory. The software uses evolutionary search to determine relationships that describe data, with the use of mathematical equations. This powerful tool allows modeling of time series. Eureqa, has also a Python API interface we can make use of for implementing our own measurement, and plotting. Weka was also used to measure the errors, and build some experiments using Linear Regression.

The GEA algorithm will make use of the Time Sliding Window approach, that aims to give a range of lags, lets name them k. As we mentioned before the selection of the parameter k can change the performance and the search space, for the learning. To experiment more with our results, we will assign different values for k and compare the performance, and also detect if some of the input variables are more influential when the sliding window changes. This analysis will take place in the fourth Chapter of the final Report. In our implementation of the GEA we will make use of the Sliding Time Window technique, that as we described in the TSF, defines the time lags and allows to build a forecast. The window size is important and for comparison reasons, we will set both the Neural Network window size parameters, as well as the, GEA on the same size. The size, can limit or overextend the search space of the model depending on large it is. The selection of the window size in our case will be the range between $t-1$ and $t-12$, which will allow to detect seasonal trends, as suggested by Cortez [23].

Regarding the Recurrent Neural Networks [20], we will develop an LSTM model for multivariate time series forecasting in the Keras deep learning library, and make use of the sklearn libraries, for encoding, defining the sequence, dealing with outliers etc. We will also perform training On Multiple Lag Time steps, on our data, and compare the results to the next machine learning method we will use, which is GEA.

3.2.2 Specifications for the Evaluation Framework

In order to evaluate the predictive performance of our machine learning models, we need to make use of the metrics that were mentioned in the Research Chapter. We will divide the Results chapter to four sections.

We will first demonstrate the results of the GEA with the help of Eureqa, in terms of how all the generated algorithms perform for the metrics included in the tables. Later on, we increase the validation set to a month each time and we measure the performance of each model, including the actual number of arrivals, and comparing that to the model's predictions. We will then compare the performance on predictions, such as to indicate the most successful models. To decide whether a prediction is accurate or not, we define a range for the percentage of prediction that was covered, and if the prediction falls under that range, then the model is considered a *winner*.

In the next section we will demonstrate the performance of RNNs on the same dataset. And also perform two independent runs, with data until the same historical points. We shall then use the RNN, to make predictions on a validation set, consisting of historic points up to the latest data, that was collected.

The overall performance of our models will be measured, by the following metrics such as the SSE, RMSE and the NMSE, but we might as well introduce more statistical methods later on in our research. To detect over fitting of our model we will then analyze our results by making use of the Bayesian Information Criterion BIC, and also the AIC and MICE, indicated to perform well for missing values. To conclude, we will present some tables of comparison between the two methods and highlight the most important observations, that we will then expand on the Conclussions chapter.

3.3 Detection of Influential Variables

Particularly, the information managers, working on the practical aspect of the arrivals in Somalia, requested to make an interpretable model of the results, including the models derived from this research. In this section we will demonstrate our experimental approach on detecting the most influential variables, meaning the variables that affect arrivals of IDPs in the state of Banadir.

The extracted regression models, from our GEA method, model the relationships between our target arrivals variable and, in all the cases, more than one predictor variables. We would like to examine how changes in the predictor values are associated with changes in the response value. In order to assist the Information Manager, a more visual and interactive representation, could better show the effects of the predictor values in the response, a simulation website.

To answer, as well the research question on the most influential variables associated with the arrivals in the region of Banadir, we need to define, what we mean with the term influential. Our definition of influential, is associated with the area of our research which is migration predictions, and the goal of the research which is, predicting arrivals. Furthermore, the methods you use to collect and measure your data can affect the seeming importance of the independent variables.

One major observation is that we should not associate the coefficients that appear in the models, with the importance of the variable they are paired with. The regular regression coefficients that we see in our models, describe the relationship between the independent variables and the dependent variable. The coefficient value represents the mean change of the dependent variable given a one-unit shift in an independent variable. Consequently, you might think you can use the absolute sizes of the coefficients to identify the most important variable. After all, a larger coefficient signifies a greater change in the mean of the independent variable. However, the independent variables can have dramatically different types of units, which make comparing the coefficients meaningless. For example, the meaning of a one-unit change differs considerably when the variables used measure money, lives, or river levels. Larger coefficients dont necessarily represent more important independent variables.

To deal with the variation of the coefficients, we can base the significance of each independent variable that appears in the model, on the sensitivity criterion. The **sensitivity** can be defined as the relative impact within a model, that a variable has on the target variable. The impact can be either a positive or a negative on the estimation of the target. Positive sensitivity of a variable means that the variable leads to an increase of units in the target variable. Negative sensitivity of a variable mean that the variable leads to a decrease of units in the target variable. Let us, try to define these notions in a more mathematical a definition of sensitive as was also described in the paper by Hosman [28], under the section *Basis for sensitivity formulas*.

Given a model equation of the form $d = f(x_1, x_2....x_n)$, the influence metrics of x_1, for example, on d are sensitivity at all data instances, is defined as follows:

$$Sensitivity = |\frac{\vartheta d}{\vartheta x_1}| * (\frac{\sigma(x_1)}{\sigma(d)})$$

where, $\frac{\vartheta d}{\vartheta x_1}$ is the partial derivative of d with respect to x_1, σ is the standard deviation of σx_1 in the input data, is the standard deviation of d.

In the next chapter, under the subsection of the most influential variables, we will analyze the sensitivity of the variables for each model and determine, the *winners*, which we will then test with both the NN and GEA approach. These variables are then given to experts, for them to interpret if and how these variables in their knowledge are associated with arrivals in the state of Banadir. We will produce a new model give the most influential variables and re-evaluate that model for predictions.

Chapter 4

Results and Evaluation

To collect the results and make comparisons, as well as make conclusions on predictions of arrivals in the state of Banadir, in Somalia, we designed three types of forecasting test.

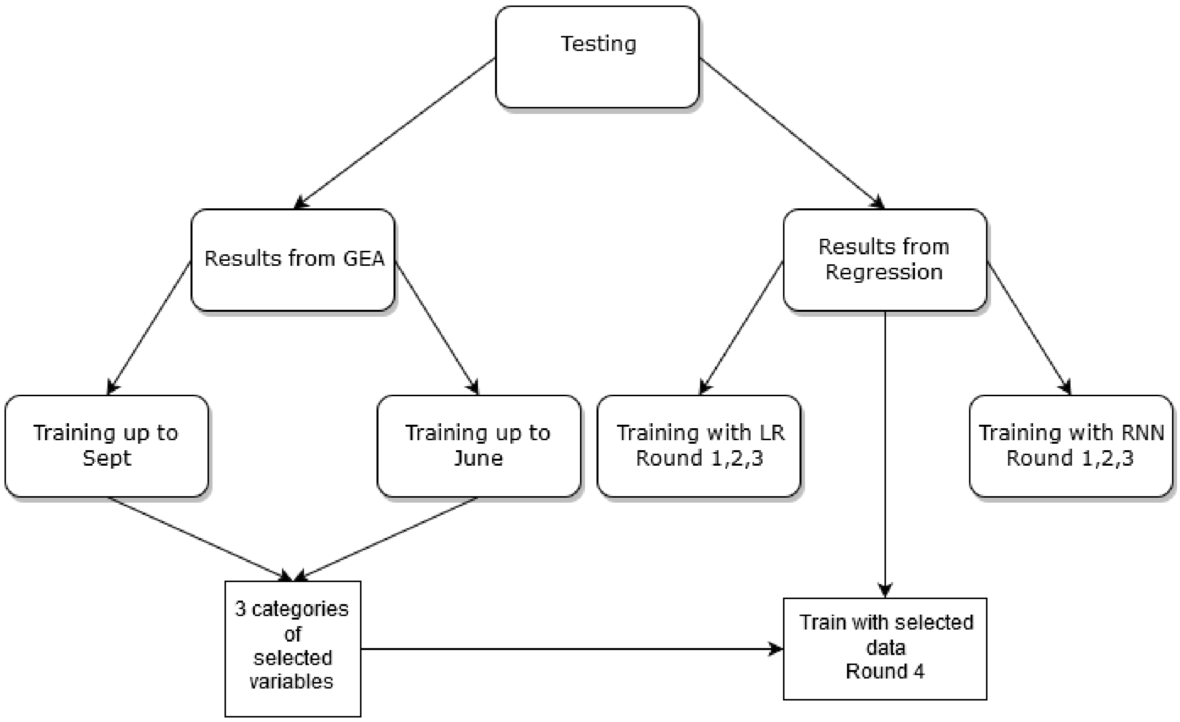

Figure 4.1: The structure of the Results

One is the generations of models, using GEA, with two different training sets, one with data up to June and one with data up to September. We will then perform a comparison of the results on the testing set, to compare the performance of the models.The second one is to do fine tune configuration with different input parameters based on our regression methods, see table 4.1.

After that, we compare linear regression with the recurrent neural networks, algo-

round	description	attributes	min lag	max lag	overlay data
1	No use of overlay data	Banadir CurrentRegion	1	12	none
2	Use all attributes	all	1	2	use and select all
3	Use all attributes max 6 lag	all	1	6	use and select all
4	Use attributes from GEA	a list*	1	max	top GEA

Table 4.1: Separating rounds of testing

rithms based on the same settings of configuration. The most influential variables, of the dataset, of the model with the best performance in the first experiment, (GEA) will be used as the input dataset on the third part. Then according to the analysis of the outcome metrics, we can filter out the optimal model.

4.1 Results of GEA

In this section we will present in the two following tables 4.2, and 4.3, the results of the evaluation of the ten models per testing data, on the validation data of the seven upcoming months. The results of the predictions and the recorded actual arrivals, of months, that were not included on the training set, are included in the next tables where we can see the numbers and how accurate each model is, see 4.4, 4.5.

metric:	Rˆ2	Correlation Coefficient	Max Error	MSE	MAE
BA1	-1.53	-0.51	101,115	2,187,880,000	35,013
BA2	-1.08	-0.10	94,736	1,797,370,000	30,005
BA3	-0.93	-0.50	94,619	1,664,130,000	26,946
BA4	-30.68	0.49	361,555	27,698,900,000	123,452
BA5	-0.85	0.03	97,268	1,619,820,000	25,976
BA6	-0.10	0.44	76,906	965,465,000	19,148
BA7	-0.53	0.23	85,785	1,333,640,000	26,004
BA8	-1.48	-0.06	100,328	2,079,440,000	29,981
BA9	-0.52	-0.09	90,256	1,333,120,000	23,126
BA10	-0.98	-0.03	96,355	1,734,690,000	28,927

Table 4.2: September / Test set up to May

metric:	R^2	Correlation Coefficient	Max Error	MSE	MAE
BAJUN1	-2.50	-0.44	35,177	446,988,000	18,372
BAJUN2	-2.73	-0.17	31,278	418,133,000	18,269
BAJUN3	-11.26	-0.12	69,751	1,567,450,000	31,887
BAJUN4	-11.26	-0.12	69,751	1,567,450,000	31,887
BAJUN5	-1.88	-0.27	27,990	371,551,000	17,152
BAJUN6	-2.89	0.02	36,101	668,213,000	22,194
BAJUN7	-3.05	-0.15	34,811	454,127,000	18,199
BAJUN8	-3.45	-0.13	41,146	498,652,000	18,874
BAJUN9	-1.56	-0.33	30,372	326,643,000	16,089
BAJUN10	-185.09	-0.14	274,802	23,798,500,000	130,301

Table 4.3: June/ Test set up to February

DATE	OCT 17	NOV 17	DEC 17	JAN 18	FEB 18	MAR 18	APR 18
REC	18461	24302	44009	44926	36822	115474	47045
BA1	42256	27093	16102	19701	15268	15191	14388
Acc	229%	111%	37%	44%	41%	13%	31%
BA2	22796	24105	20941	17872	19536	20994	69900
Acc	123%	99%	48%	40%	53%	18%	149%
BA3	24664	29302	119842	33099	22546	20702	20745
Acc	134%	121%	272%	74%	61%	18%	44%
BA4	9172	49619	105586	228676	145010	241011	411405
Acc	50%	204%	240%	509%	394%	209%	874%
BA5	14149	19780	16655	16528	16824	18224	36108
Acc	77%	81%	38%	37%	46%	16%	77%
BA6	17654	28071	25106	23955	30165	38568	54229
Acc	96%	116%	57%	53%	82%	33%	115%
BA7	31357	23118	25883	22219	23380	29689	17629
Acc	170%	95%	59%	49%	63%	26%	37%
BA8	20343	24358	115168	33859	25803	16990	10797
Acc	110%	100%	262%	75%	70%	15%	23%
BA9	25026	27022	30162	28806	28004	25218	22031
Acc	136%	111%	69%	64%	76%	22%	47%
BA10	19329	19844	17246	14324	18197	19119	76560
Acc	105%	82%	39%	32%	49%	17%	163%
MEAN	22675	27231	49269	43904	34473	44571	73379
Acc	123%	112%	112%	98%	94%	39%	156%

Table 4.4: Predictions for Banadir after September

DATE	JUL 2017	AUG 2017	SEP 2017	OCT 2017	NOV 2017	DEC 2017
REC	39219	25768	21554	18461	24302	44009
BA_JUN1	17818	7984	57002	39243	42373	27847
Acc	45%	31%	264%	213%	174%	63%
BA_JUN2	19244	12023	54865	39794	50267	31931
Acc	49%	47%	255%	216%	207%	73%
BA_JUN3	10869	19227	35864	64746	93937	72711
Acc	28%	75%	166%	351%	387%	165%
BA_JUN4	21100	-28455	68036	39955	45194	40472
Acc	54%	-110%	316%	216%	186%	92%
BA_JUN5	15639	-11974	48997	39191	41049	26906
Acc	40%	-46%	227%	212%	169%	61%
BA_JUN6	25033	15476	57761	46114	59480	30833
Acc	64%	60%	268%	250%	245%	70%
BA_JUN7	21983	8322	55994	22040	44997	34519
Acc	56%	32%	260%	119%	185%	78%
BA_JUN8	20580	11621	62279	41227	44986	32642
Acc	52%	45%	289%	223%	185%	74%
BA_JUN9	18133	9773	52278	36192	40534	32233
Acc	46%	38%	243%	196%	167%	73%
BA_JUN10	14391	12937	90981	149407	296042	199175
Acc	37%	50%	422%	809%	1218%	453%
BA_TOTAL	18479	5693	58406	51791	75886	52927
Acc	47%	22%	271%	281%	312%	120%

Table 4.5: Predictions for Banadir after June

4.2 Results of Regression Algorithms

The data set is categorized as a series of data representing a specific variable for arrivals in the state of Banadir, as well as data combination of the whole entity of Somalia, indicating internal migrations, as well as other indicators. Based on the properties of our structure, the data set is a time series structure, where the regions' indicators are independent attributes and the Banadir_CurrentRegion is dependent attributes.

The task of data pre-processing, since we will be executing our algorithms in python and Weka, in this section is different than the process in the previous section. In this section, we need to do the following 2 pro-process:

- Transform the matrix of variables to include all the time lags to include in the training set.
- Include the time-stamp in a format that allows the time-stamp to be included as a power of time.

4.2.1 Round 1 to Round 3

For each Round, we will present the statistics collected for each step of prediction. The steps represent the months in the first column of the three tables below 4.6,4.7,4.8. The highlighted cells, are indicating whether we detected an extreme error, or an error that is acceptable.

round 1	rec	LR prediction	NN predicted	LR error	NN error
Jul , 2017	39219	84827.8581	118295.3427	45608.85	79076.34
Aug , 2017	25768	4394.637	92971.9551	18163.6	67203.95
Sep , 2017	21554	-23115.3472	32891.3033	-44670.65	11337.30
Oct , 2017	18461	53887.8158	184110.7927	35426.81	165649.79
Nov , 2017	24302	50074.1711	113472.2235	25772.17	89170.22
Dec , 2017	44009	52909.1307	119991.2688	8900.13	75982.26
Jan , 2018	44926	6311.9915	102375.9379	-38614	57449.93

Table 4.6: Round 1: The statistics for 7 months predictions

The statistics data of Mean absolute error (MAE) for 3 round testing is shown in table below 4.9.

The graph 4.2 can show us in more clear way that 3-LR and 2-NN is performing better than the rest.

The statistics data of Root mean squared error (RMSE) for the three rounds of testing is shown in 4.10.

round 2	actual	LR prediction	NN predicted	LR error	NN error
Jul , 2017	39219	7968.2973	26942.2376	-31250.7027	-12276.7624
Aug , 2017	25768	18648.9503	39740.0176	-7119.0497	13972.0176
Sep , 2017	21554	23219.6201	28715.4903	1665.6201	7161.4903
Oct , 2017	18461	31834.0332	59268.7311	13373.0332	40807.7311
Nov , 2017	24302	4356.253	29756.928	19177.253	5454.928
Dec , 2017	44009	7968.9709	6233.9953	-36040.0291	-37775.0047
Jan , 2018	44926	45533.4611	29894.7413	607.4611	-15031.2587

Table 4.7: Round 2: The statistics for 7 months predictions

round 3	actual	LR prediction	NN predicted	LR error	NN error
Jul , 2017	39219	31587.3535	30900.6268	-7631.6465	-8318.3732
Aug , 2017	25768	41833.2536	27986.1455	16065.2536	2218.1455
Sep , 2017	21554	14631.5535	30948.7038	-6922.4465	9394.7038
Oct , 2017	18461	38448.8034	1466.1383	19987.8034	-16994.8617
Nov , 2017	24302	46798.4641	1422.2377	22496.4641	-22879.7623
Dec , 2017	44009	9041.7299	1421.6213	-34967.2701	-42587.3787
Jan , 2018	44926	15922.312	1435.6043	-29003.688	-43490.3957

Table 4.8: Round 3: The statistics for 7 months predictions

round	1 - LR	1 - NN	2 - LR	2 - NN	3 - LR	3 - NN
1-step	47619	76208	25899	25397	21186	32075
2-step	47842	108070	25331	28374	22636	34554
3-step	51552	134305	27665	30786	23466	38387
4-step	52535	187487	31451	34215	25880	42393
5-step	72799	226638	34483	32872	26624	46573
6-step	76980	236692	37538	37520	27413	51078
7-step	99858	207254	37973	37710	26006	52996
8-step	99127	191799	50340	45049	24838	56164
9-step	115389	172712	65042	58818	36537	69189
10-step	49193	163641	15654	16810	11360	45593

Table 4.9: 10 months of predictions Mean absolute errors

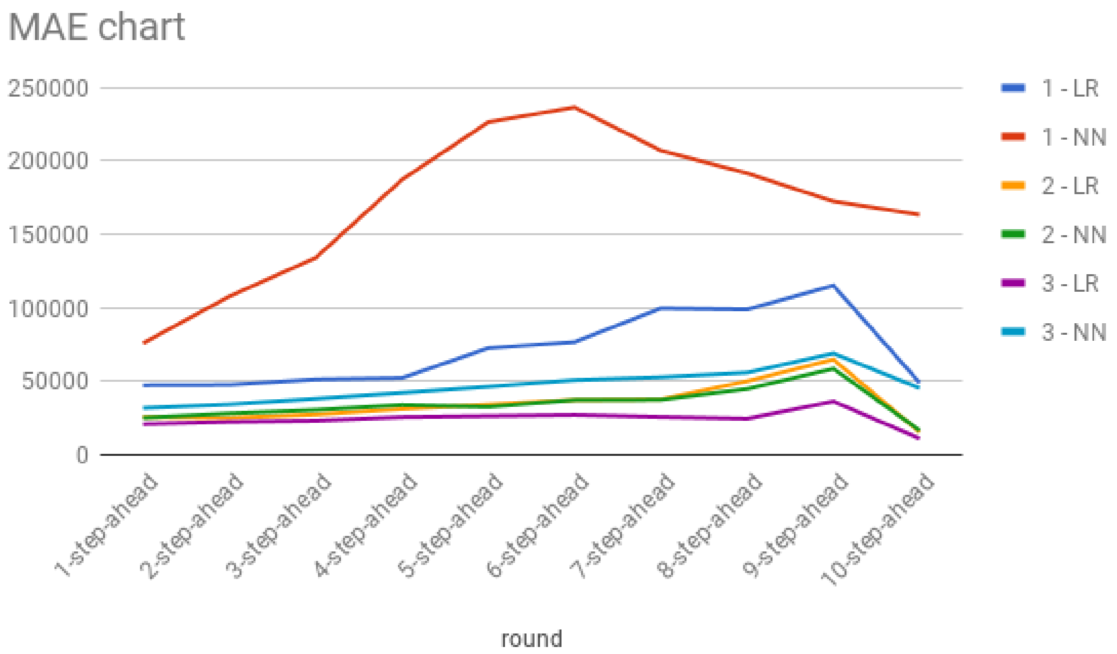

Figure 4.2: MAE Comparison of Rounds

round	1 - LR	1 - NN	2 - LR	2 - NN	3 - LR	3 - NN
1-step	55426.10	84802.18	40746.91	37343.89	26943.55	41664.16
2-step	56411.54	118949.50	41690.50	39670.13	28267.73	43562.86
3-step	59487.86	148587.96	44168.53	42239.95	29444.53	45886.87
4-step	61312.98	206954.65	47216.77	45096.81	31554.13	48709.12
5-step	79423.22	236203.81	50709.83	45548.83	32755.94	52077.47
6-step	83850.91	238380.82	54881.54	49705.59	34479.39	55567.47
7-step	107455.06	207635.98	58691.01	52444.33	34679.61	58064.34
8-step	111396.20	194442.67	67768.66	59875.46	36235.37	62168.56
9-step	133028.42	176306.91	81668.56	72279.48	44370.03	73102.45
10-step	49193.57	163641.17	15654.60	16810.72	11360.84	45593.31

Table 4.10: 10 months of predictions RMSE

The graph 4.3 can show us in more clear way that 3-LR and 2-NN is performing better than the rest.

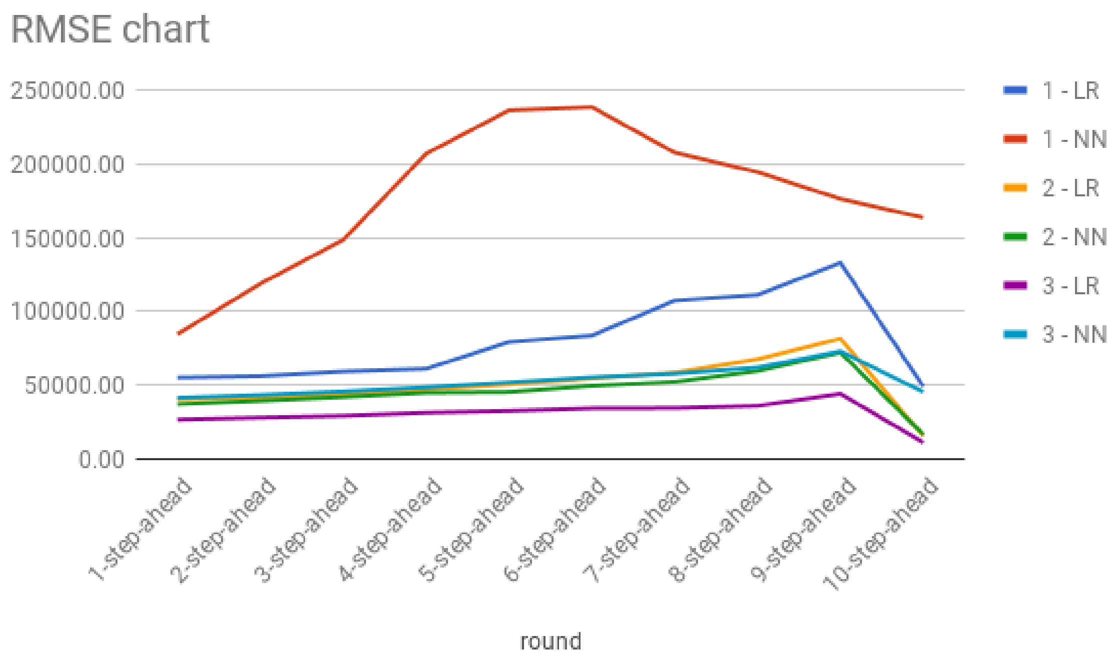

Figure 4.3: Comparison of Rounds

Both metrics, indicate that the round with the best average performance of prediction is Round three, and also the 2-NN is worth revising. Besides, we can find a significant error reduction from round 1 to round 4, meaning that somehow the independent variables affect the prediction in a positive manner. In the next section we will perform variable evaluation to select the dataset, for Round four.

4.3 Most Influential Variables in GEA

In this section, we will use the sensitivity metric, that was introduced in Chapter 3, to detect the variables that affect the outcome of the models produced by genetic algorithms. We will first show which independent variables, have a high ranking, what is the frequency of these variables, in the models, in two separate tables, where each table represents the models with training sets up to June and up to September. The tables only represent a sample and the rest of the variables can be found in the appendix, along with all the moles.

In order to reduce the independent variables, we are going to use in the next step of our experimental set-up, we will detect the common variables between the

Variable GROUPED	Sensitivity	Frequency
Gedo_rain	20.75	3
Hiiraan_BeforeRegion	11.50	9
Bay_BeforeRegion	7.35	10
Banadir_BeforeRegion	3.38	10
Mudug_Fatalities	2.59	8
Nugaal_CurrentRegion	1.44	4
Mudug_BeforeRegion	1.37	8
Gedo_BeforeRegion	1.22	3
Bari_Conflict	1.04	7
Sool_CurrentRegion	0.95	2
Hiiraan_WaterDrumPrice	0.84	5
Awdal_Conflict	0.72	2
Shab_D_BeforeRegion	0.52	5
Nugaal_FutureRegion	0.36	4

Table 4.11: Top most influential variables on June data

Variable GROUPED	Sensitivity	Frequency
Bakool_WaterDrumPrice	9.54	9
Mudug_BeforeRegion	5.00	10
Mudug_FutureRegion	2.63	2
Hiiraan_BeforeRegion	2.55	4
Gedo_CurrentRegion	1.85	9
Bari_FutureRegion	1.46	2
Galgaduud_FutureRegion	1.36	4
Banadir_WaterDrumPrice	1.31	1
Togdheer_CurrentRegion	1.29	1
Bay_BeforeRegion	1.29	2
Bari_CurrentRegion	1.11	4
Jubbada_Hoose_goatprice	1.06	5
Togdheer_BeforeRegion	0.97	2
Sanaag_Fatalities	0.85	4
Shab_D_Fatalities	0.78	2
Gedo_BeforeRegion	0.67	4
Bakool_rain	0.64	1
Jub_D_CurrentRegion	0.61	3
Gedo_goatprice	0.40	2

Table 4.12: Most influential variables on September data

two tables shown above4.12,4.11. These common variables that possibly affect migration, we shall discuss further in the next chapter.

Hiiraan_CurrentRegion	Bari_Conflict	Gedo_FutureRegion	Gal_FutureRegion
Sanaag_CurrentRegion	Bakool_WaterPrice	Awdal_Conflict	Shab_D_Fatalities
Gedo_Conflict	Mudug_Fatalities	Juba_River_discharge	Mudug_BeforeRegion
WG_FutureRegion	Sool_CurrentRegion	WG_rain	Tog_CurrentRegion
Jub_H_BeforeRegion	Gal_BeforeRegion	Gedo_Fatalities	Tog_Fatalities
BW_Shab_River	BB_Shab_River	Nugaal_FutureRegion	Awdal_BeforeRegion
Bakool_rain	Nugaal_Fatalities	Bakool_FutureRegion	Bay_BeforeRegion
Jub_D_CurrentRegion	Galg_WaterPrice	Hiir_WaterPrice	Tog_BeforeRegion
Nugaal_WaterPrice	Tog_WaterPrice	Mudug_CurrentRegion	Shab_H_rain
Nugaal_Conflict	Gal_CurrentRegion	Bari_Fatalities	Gedo_BeforeRegion
Hiir_BeforeRegion	Jub_D_Conflict	Bak_CurrentRegion	San_Fatalities
Bari_CurrentRegion	Jub_D_FutureRegion	Shab_D_rain	Bari_rain
Mudug_FutureRegion	Ban_WaterPrice	Bari_FutureRegion	Sanaag_Conflict
Hiir_Conflict	Gedo_goatprice	Jub_H_FutureRegion	SD_BeforeRegion
Nugaal_CurrentRegion	Gedo_CurrentRegion	Jub_H_goatprice	D_Juba_River
Banadir_BeforeRegion	Bay_WaterPrice	Gedo_rain	

Table 4.13: Merged variables appearing in both groups

Gedo_FutureRegion	Shab_D_Fatalities	Nugaal_Conflict	Bari_rain
Awdal_Conflict	Mudug_BeforeRegion	Hiiraan_BeforeRegion	Sanaag_Conflict
Gedo_Fatalities	Tog_CurrentRegion	Bari_CurrentRegion	S_D_BeforeRegion
Nugaal_FutureRegion	Togdheer_Fatalities	Mudug_FutureRegion	Bakool_WaterPrice
Hiiraan_WaterPrice	Awdal_BeforeRegion	Nugaal_CurrentRegion	Mudug_Fatalities
Mudug_CurrentRegion	Bay_BeforeRegion	Banadir_BeforeRegion	Sool_CurrentRegion
Bari_FutureRegion	Tog_BeforeRegion	Gedo_rain	Gal_BeforeRegion
Jub_H_goatprice	Bari_Conflict	Gal_FutureRegion	Gal_CurrentRegion
Bakool_rain	Gedo_BeforeRegion	Gedo_CurrentRegion	Banadir_WaterPrice
Jub_D_CurrentRegion	Sanaag_Fatalities	Gedo_goatprice	

Table 4.14: Merged variables with high sensitivity appearing in both groups

4.4 LR and NN with reduced dataset

In this section of the chapter, we will make use of the results of the previous sections, to apply LR and NN to a reduced dataset, with different parameter tuning as we can see in the table 4.16.

round	description	attributes	min lag	max lag	overlay data
4.1 LR 2	LR on common vars	4.13	1	2	all
4.1 NN 2	NN on common vars	4.13	1	2	all
4.1 LR 6	LR on common vars	4.13	1	6	all
4.1 NN 6	NN on common vars	4.13	1	6	all
4.2 LR 2	LR of top common vars	4.14	1	2	all
4.2 NN 2	NN of top common vars	4.14	1	2	all
4.2 LR 6	LR of top common vars	4.14	1	6	all
4.2 NN 6	NN of top common vars	4.13	1	6	all

Table 4.15: Separating runs for Round 4

Running the above experiments we collected the metrics of each run, and that gives us the table below. According to which the line chart is created, but since there is a relatively huge difference in the errors, the graph is used to show the comparison of the rest, while the 4.1 NN 6 and the 4.2 NN 6 is excluded from the chart.

round	4.1LR2	4.1NN2	4.1LR6	4.1NN6	4.2LR2	4.2NN2	4.2LR6	4.2NN6
1-step	27938	31852	27217	554612	29791	23482	66367	480642
2-step	29089	36178	29116	3382389	30073	24778	69290	2341089
3-step	26531	35246	31520	6043608	32107	25186	68984	3522987
4-step	30207	36231	31619	7419714	35324	27798	59943	4025078
5-step	32453	39885	34913	8656384	40129	31720	57168	4690891
6-step	31697	44137	41882	10381294	45688	35066	62763	5717919
7-step	30795	44767	41847	12938103	49100	35038	63546	8695947
8-step	35793	47640	41960	17334135	54480	36218	60187	9243656
9-step	36492	59330	56219	17350346	69292	41215	75526	9213622
10-step	6509	18860	1741	17384561	35078	27690	4060	9246936

Table 4.16: MAE for round 4 subruns

We can therefore, see clearly that the performance of the Round 4.2 of Neural Networks with a time lag of 2 of the top merged variables, as well as the Linear Regression with a lag of 2 of the merged variables, perform the best out of all the models. Let us see in detail the results of the predictions and compare them to the results of the ten recorded months.

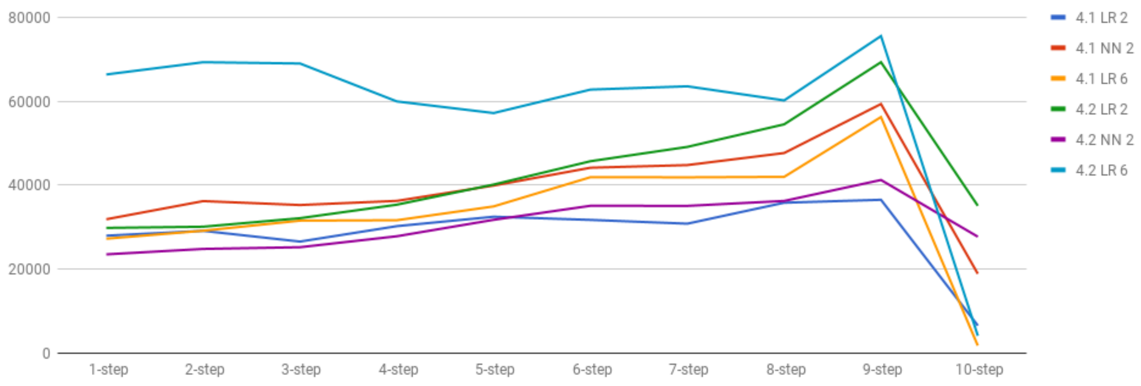

Figure 4.4: Comparison of MAE Round 4

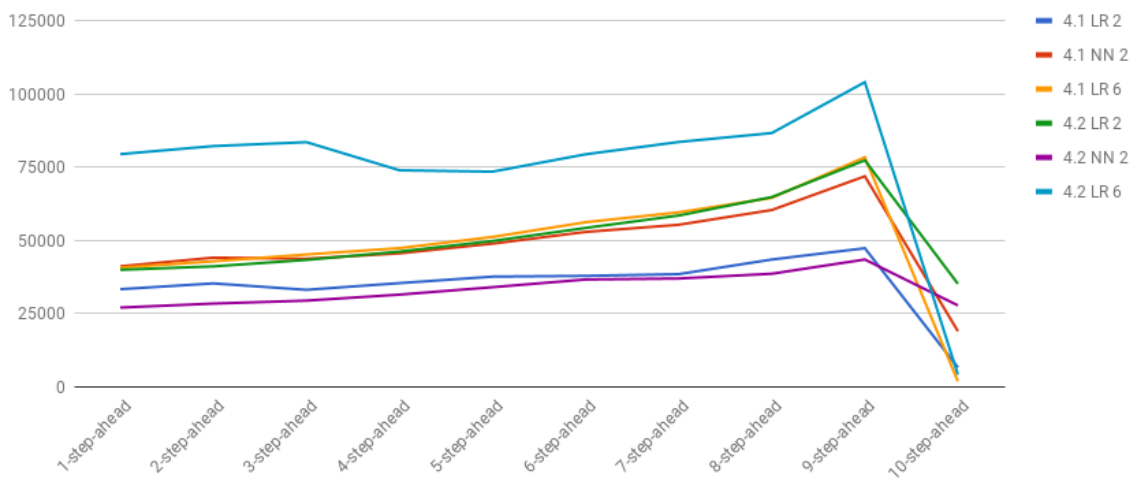

Figure 4.5: Comparison of RMSE Round 4

step	actual	NN lag 2	match	LR lag 2	match
Jul , 2017	39219	18099.92	46.15%	35309.09	90.03%
Aug , 2017	25768	60565.36	235.04%	74775.73	290.19%
Sep , 2017	21554	29597.58	137.32%	28366.36	131.61%
Oct , 2017	18461	19540.82	105.85%	34377.83	186.22%
Nov , 2017	24302	9044.29	37.22%	-10045.2	-41.33%
Dec , 2017	44009	8798.25	19.99%	4191.68	9.52%
Jan , 2018	44926	16874.53	37.56%	25945.38	57.75%
Jul , 2018	36822	14529.63	39.46%	70270.81	190.84%
Aug , 2018	115474	66681.81	57.75%	53317.3	46.17%
Sep , 2018	47045	26868.81	57.11%	32059.02	68.15%

Table 4.17: Predictions of the top performing rounds

Chapter 5

Conclusions and recommendations

In this chapter, is presented a detailed extraction of conclusions, alongside with different approaches to verify those conclusions, such as extra experiments and experts' opinions.

5.1 Conclusions

Normally, in the conclusions chapter, a response is given to each of the research questions. We shall begin this chapter by observations made since the beginning of the project, during the collection of data until the last part of the projects, during the set up of the experiments. In the end we will make assumptions based on the results, collected and to complete the chapter we will proceed with recommendations.

The first phase, of this project, was the definition of the project, determining what the goal should be and what is the background of the field that would be explored. The worldwide phenomenon of migration, was an extremely sensitive topic to tackle, even more, in the Horn of Africa. In the family of countries, located in the Horn of Africa, the situation is unsteady. Autonomous military groups, terrorism, conflict based on religion are some examples that intensify the unbalanced situation. The country of Somalia, was the region of interest, and the positive aspect of limiting the scope to one country that had availability on data collections, since the country has been under continuous internal displacement, for a long period of time, approximately eight (8) years. The data officers, as well as some national organizations were creating reports including tables, graphs, text containing numbers, that would reflect all the incidents and movement, detected within the country. Other sources were collecting *numbers* of interest, such as economic factors and climatological factors. Specifically, one data collection that was reflecting the reports collected in the camps, which then combined was projecting the exact number of arrivals, in each of the states. These data sources, sparked the interest of the information officers to request for the unknown yet, expectancy of arrivals in the state they corresponded to. The existing data and the request formated the basis of research for this topic. *What can we do to predict arrivals to a state?*

Once the scope of interest was limited to predicting arrivals for one of the states regions of Somalia, then some research was required to guide the project towards a possible solution to the problem. Throughout all the literature research, which included independent articles, discussions with experts and statisticians, the phenomenon of migration started to form a shape. A lot of effort, has been made in the past to explain models that can describe migration and to extract the driving factors of displacement. Concluding, from the theoretical research, a lot of statistical methods have been used to explain the phenomenon, and most of the models describe migration to be explicitly dependent on distance and prospect in the region of destination. Our case, of internal displacement differs, in a way that it does not refer to a cross-border migration, and it clearly relates to survival, in the region of destination. Another important aspect of this internal displacement, is also the temporarily of displacement. People migrate due to climate conditions, in the more prosper regions, intending to return as soon as their region of origin becomes suitable for their families, as was reported by experts. The mathematical models to describe migration either failed or could not be further explained due to complexity.

Meanwhile, the field of machine learning has been using forecasting in many fields, to make predictions, commonly, regarding the stock market, energy and product consumptions, and much more. The many applications of machine learning in fields as such, can very accurately give estimations and model a set of input variables to either perform classification or regression, to extrapolating future values. In our case, a few cases of forecasting on time series, has been explored with machine learning, and much less in the humanitarian field, for migration. To guide the methodology used for this thesis, the base was the approach in different fields, which also try to detect the driving changes in the data that affect the target variable. The approach we took, due to existing research and also the interpret-ability of the models, was to explore our problem by applying two different methods, one of which being Neural Networks and the other one Genetic and Evolutionary Algorithms.

In the next paragraphs we will explain the results and the conclusion of the experiments, but first we need to reflect on the data. The data, as was described in the previous chapter, consisted of a variety of data collections, including economic factors, climate indicators, tracking of movement, and incidents, conflicts and deaths. To collect seven (7) years of data, we had to parse data in all sorts of formats such as pdf, excel, graphs, reports and thus this project resulted in many assisting scripts and crawlers, as well as the use of external software, to serve the many needs of parsing. Throughout all this process, plenty of data sources were rejected due to discontinuity, lack of trust in the source, as well as conflict of data upon cross-validation. In the Recommendations section we will include some useful guidelines for future researchers.

Focusing on our experimental set-up we used our data collection as an input for the algorithms, and for the actual implementation of the algorithms, with the use of either programming languages, such as python or software solutions developed for academic research, such as Eureqa and Weka. These powerful tools, resulted in making predictions as we showcased in the Results chapter. Our conclusions on the results, in the same order as they were presented in the last chapter, will be described in the section below:

1. **Genetic and Evolutionary Algorithms**
   Concerning the first trial of training the machine, on the dataset with training data until June, we collected inaccurate predictions for the seven (7) upcoming months, which was our validation set. While the fit of the models, on the training data was good and the mean absolute error as we can see in the table 5.1, are comparatively low, the model failed to predict future values of arrivals reaching 312% of accuracy on November, extremely wrong levels. We will consider this case of failure an over-fitting of the model to the training set. We believe that the models for June, learned to interpret the noise in our data set to the extend that it impacted the future predictions in a negative way. The concepts that the models were trying to fit, in order to decrease the error metrics, made them unable to generalize and apply to the validation set. For the curiosity of the reader, in the Appendix section, we will include more of the metrics on the training and the testing sets. RMSE and MAE, were selected to give high penalty to large errors and an overview of the errors on each set, accordingly.

   2. Regarding the dataset with training data until September, the predictions collected are considered accurate and some even precise to the actual numbers of arrivals in the region of Banadir. The first five months of predictions reached precision from 94% to 123%, while it failed for the last two months of the validation set. March and April, reached 39% and 156% of accuracy, accordingly, which indicates that the models no longer were describing the phenomenon of arrivals in Somalia in a fitting way. In order to validate our theory and to investigate further, how the genetic and evolutionary algorithms perform when fed with more data, we created a dataset with data up to January and here are the results. Thus, we generated ten more models fed with data up to January of 2018 and we shall compare the predictions in the table 5.2 below.

   What we observe in the table above? Is that we cannot say with certainty that

Metric	Train ME	Test ME	Train MAE	Test MAE	Train RMSE	Test RMSE
BAJUN1	22639	35177	2245	18372	4179	21142
BAJUN2	14818	31278	2237	18269	3541	20448
BAJUN3	10549	69751	2057	31887	3203	39591
BAJUN4	10549	69751	2057	31887	3203	39591
BAJUN5	14572	27990	2309	17152	3497	19276
BAJUN6	21379	36101	2511	22194	4385	25850
BAJUN7	8326	34811	1935	18199	2644	21310
BAJUN8	11495	41146	2363	18874	3620	22331
BAJUN9	14041	30372	2319	16089	3594	18073
BAJUN10	20927	274802	2381	130301	4284	154268

Table 5.1: Comparison of Metrics with Test and Train set

DATE	Feb , 2018	Mar , 2018	Apr , 2018
ARRIVALS	36822	115474	47045
BA_JAN1	49528	41348	40061
Acc	134.51%	35.81%	85.16%
BA_JAN2	65416	28046	18619
Acc	177.65%	24.29%	39.58%
BA_JAN3	46107	100213	86162
Acc	125%	87%	183%
BA_JAN4	46350	41518	59766
Acc	126%	36%	127%
BA_JAN5	53218	31257	24279
Acc	145%	27%	52%
BA_JAN6	17707	22280	78599
Acc	48%	19%	167%
BA_JAN7	54032	56178	63449
Acc	147%	49%	135%
BA_JAN8	36047	50762	71130
Acc	98%	44%	151%
BA_JAN9	58464	34254	26499
Acc	159%	30%	56%
BA_JAN10	69454	39847	83633
Acc	189%	35%	178%
TOTAL AvG	49632	44570	55220
Acc	135%	39%	117%

Table 5.2: Predictions for Banadir after January

we can predict arrivals in the region of Banadir based on Genetic and Evolutionary Algorithms. But what we can see is that the model BA_JAN3 managed to predict the pick in arrivals for March, which was an extreme number of arrivals for the region of Banadir. That matching number, indicates that some of the factors present in that model did affect people to move in the region, and that somehow these variables found a fit in the training set to describe past arrivals as well.

Taking into account the results, and going through the process of making models with the use of GEA, we noticed many models were performing well on training, but relatively bad in predicting. Of course, we cannot expect accurate predictions all the time, but we cannot also trust the results of the models to say with certainty that we can now predict arrivals in the region of Banadir.

2. **Regression Algorithms**
In order to explore the regression techniques that could lead to predictions, we decided to run different experiments, varying on maximum time lags and input variables. The first round **Round 1** of experiments, resulted in bad predictions for all the models, the worst one being prediction on univariate time series with linear regression and neural networks of at time lag of delay of twelve (12) months, so we selected the remaining ones with the best error metrics for Round 4.

   On **Round 2**, we created a window of delay up to two months, where we saw that the predictions from both Linear Regression and Neural Networks were improved, but the predictions did not result to being accurate. For a few cases the errors were decreased significantly giving some good numbers of predicted arrivals. Increasing the window of delay to six months **Round 3**, we observed that the metrics are similar to those of Round 2, with the Linear Regression for lag of six (6) being the best. Because of the small differences between the RMSE and the MAE of these rounds, we decided to replicate the experiment on Round 4 but instead of changing the window size, we would reduce the input dataset and make comparisons based on that.

   To reflect on the comparison of all the rounds, we can see that given a window of up to six months we receive more accurate predictions. More questions arise when we compare the runs with different lags. What did we conclude from the runs, with Linear Regression, why the lag of 6 is better and how results are improving when more time lag is given, does that lead us to believe that variables in the past can reflect arrival up to 6 months in the future? A final observation, is that as we were giving the models more past observations as input, the models improved but not noticeably, for us to conclude that arrivals in Banadir are affected by observations in other regions up to six months in the past. Since internal displacement is a temporary solution for the residents of a region and they plan to return to the region of origin, is does not appear logical to use a window that looks will create a model affected by data more than six months before the incident of arrivals in the region.

3. **Most Influential Variables**

   During the phase of calculating the sensitivity of the variables in our GEA algorithms, we performed analysis on the models as separate sets, one being the set of training data up to June and the other one of data up to September. A noticeable actuality, is that there are a lot of repeated variables in the models internally in each set as well as amongst them. To proceed with the analysis, the variables were cataloged and showcased in the results section. In order to comprehend more on how these variables can influence people fleeing to Banadir, we mapped them, in a visual representation5.1, and we will make an interpretation of these variables in collaboration, cross validating our conclusions with experts.

   We conclude that Fatalities, and Conflict spread throughout the country are associated with people fleeing in Banadir, while regions of origin vary. Another conclusion is that Water Drum Price in the region, as well as the neighboring regions affects people to migrate, as we can see Hiiraan and Bakool Water Drum Price, can lead to the increase in the number of arrivals as these numbers have a positive magnitude in the modeled predictions for Banadir. Another thing that was noted by experts, when presented with the graph was that if they were to compare it to the Food Security and Nutrition Analysis, they would see that the regions that are included in one way or the other on the influencers, are the same regions that are facing severe famine, as can be seen in the map online [29]. Contrary to the observed data, of arrivals in Banadir being mostly coming from Bay and Lower Shabbelle, it appears that the variables such as the market prices, or incidents associated with those regions are not affecting the models at all. Bay and Lower Shabbelle are regions that are under the influence of Al Shabaab, the biggest military group and those regions are associated with conflict, leading to people fleeing to the neighboring regions.

4. **LR and NN with reduced dataset**

   In order to explore more on the ideal inputs for the machine learning to create good models, and having as a basis the round 2 and round 3 from the Regression techniques, we run 4 more experiments, and collected the results. The results collected do not project much better accuracy or improvement. We

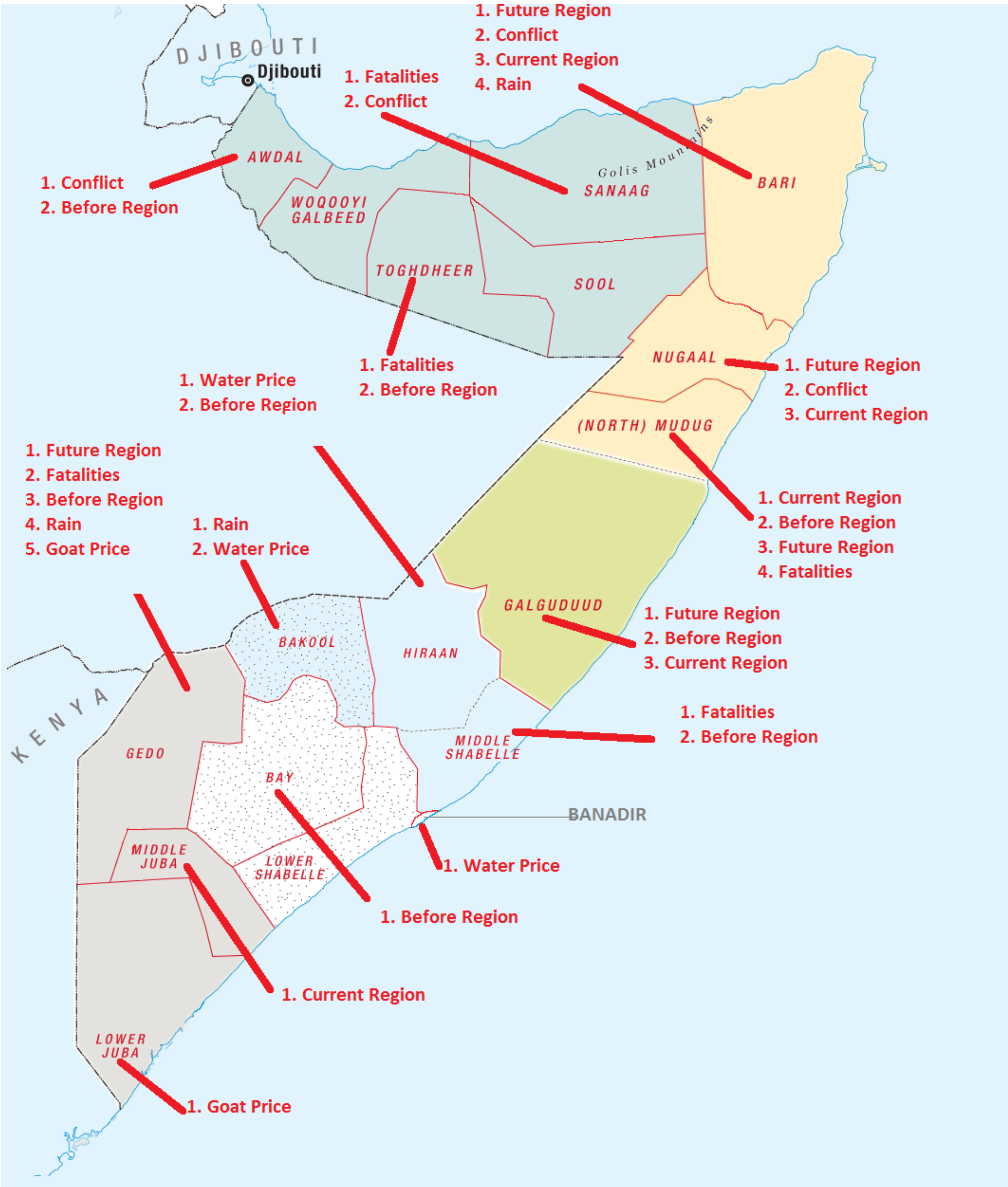

Figure 5.1: Variables by Region that affect the arrivals in Banadir

conclude from the comparison of the MAE and RMSE for round 4, that Neural Networks performs the best with a time lag of 2 units. In the next figure 5.2 we shall compare the RMSE for the winner 4.2 NN max lag 2 and the 3 LR max lag 6.

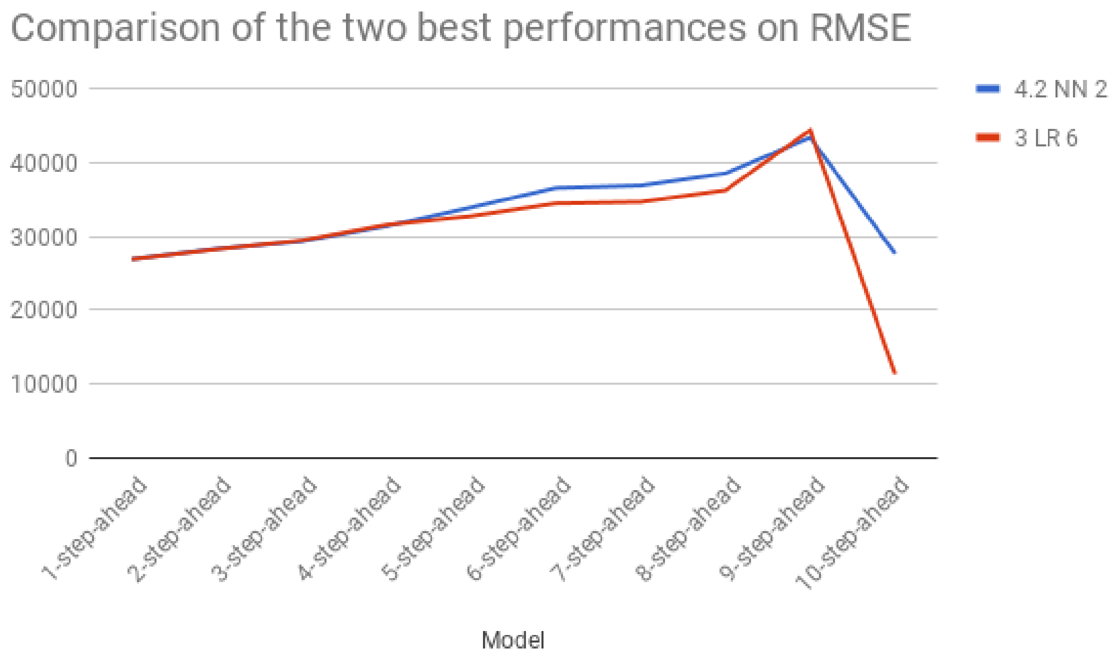

Figure 5.2: Round 4 VS Round 3 best performances

Another observation made when training the algorithms, was that there were cases of under fitting, where a base value was selected and repeated throughout all the time units in the training set. These models were excluded from the graphs due to poor performance on the training data. -

5.2 Recommendations

In order to guide future researchers, and also for this project, we will reflect on angles that we neglected due to time or data limitations, and make recommendations for further expanding this project, or to inspiring others in taking different directions. To improve this experiment, if it was to be repeated we will now make some suggestions that would change the set up either completely or enhance it, so that the results would change due either to data input or methodology.

1. Explore the seasonality of the data, such a path, would require experts to determine the different climate seasons in Somalia, in order to detect seasonal

trends and treat the data collection accordingly. Similar methodologies are followed in machine learning when there is need to predict demand of a product in the market, and for example seasons such as Christmas is excluded from the training set to remove the outliers. Another treatment of data could also be applying the technique of excluding some months from the dataset due to a temporary phenomenon, such as elections season, this would also require an expert to detect and exclude the data from the training. In product forecasting demand, holidays are usually excluded, so there could also be space for such exclusions in our dataset as well.

2. Regarding the data collection, it would be wise to pursue more data such as job openings, agricultural production in farming areas, and if allowed more data on terrorist organizations that take action, in some regions and how these incidents influence movement. Such data sources are not available in Somalia, but ideally since migration is connected with opportunistic motives, such criteria could help select the region of destination and describe the IDPs cognitive process, more accurately. Population fleeing due to extreme unsafe conditions in a region, could be pursuing a safe stable region. To classify some regions regarding safety and to which degree could be a different way to model the country of Somalia.

3. Another approach to understand the driving factors of migration, would be to perform surveys on the IDPs in furtherance of knowledge of association between our interdependent variables, so as to form an explanation of the phenomenon, and so as well to detect the maximum lag to consider for training the machine learning algorithms. Such a practice would then inform us in more detail, given that we can collect the indicated data, if the phenomenon can be modeled or it is a completely random process.

4. The research could also be expanded regarding the distance, and the role it has on migration within Somalia. As well as satellite imagery is a possibility for this project, due to the fact that already satellite imagery is used to detect catastrophes, either of natural nature or due to conflict. Since these applications already exist, it would be wise to connect them with migration and see if they could also work as indicators that people would flee from a region.

5. Another and more unorthodox way to approach this, would be to simulate the population of Somalia with multi-agent systems, and all the elements of civilization that morph the region and to allow the algorithm to show us how each agent, assuming an IDPs is an agent, would react to changing conditions of their regions. Such a simulation would require assigning penalties and weights

to roads between camps and crossing borders, but it would more accurately represent the situation. Such a system was built to simulate the tracking of movement in Nigeria by [30].

Considering this was the first attempt to model internal displacement, I believe a lot of considerations were left aside during this research, but as there was not pre-existing methodology for approaching this problem, and past theories proven to work, we had to create our own methodology according to approaches from other fields. I hope this research can be used as a basis, for further researchers, to explore the phenomenon of migration and how machine learning can serve in predicting accurate movements of population.

Bibliography

[1] D. of Economic and S. Affairs, “International migration report 2017,” United Nations, New York, 2017.

[2] OCHA, “Horn of africa humanitatian outlook,” OCHA, January 2017.

[3] S. M. Hsiang and M. Burke, “Climate, conflict, and social stability: what does the evidence say?” *Climatic Change*, vol. 123, no. 1, pp. 39–55, Mar 2014.

[4] “Web site Migration data portal The bigger picture.” [Online]. Available: https://migrationdataportal.org/

[5] GMDAC, “Migration forecasting: Beyond the limits of uncertain,” Data Briefing Series, November 2016.

[6] J. Bijak, *Forecasting Migration: Selected Models and Methods.* The Springer Series on Demographic Methods and Population Analysis, 2010.

[7] M. Kupiszewski, “The role of international migration in the modelling of population dynamics,” Ph.D. dissertation, Institute of Geography and Spatial Organisation, Polish Academy of Sciences, Warsaw, 2002.

[8] D. Kupiszewska and B. Nowok, “Comparability of statistics on international migration flows in the european union.” CEFMR Working Paper, 2005.

[9] N. Kamel, A. Atiya, N. Gayar, and H. El-Shishiny, “Tourism demand foreacsting using machine learning methods,” 05 2018.

[10] F. Kadri, F. Harrou, S. Chaabane, and C. Tahon, “Time series modelling and forecasting of emergency department overcrowding,” vol. 38, p. 107, 09 2014.

[11] F. Simini, M. C. Gonzalez, A. Maritan, and A.-L. Barabási, “A universal model for mobility and migration patterns,” *Nature*, vol. 484, pp. 96–100, 2012.

[12] M. Lenormand, A. Bassolas, and J. J. Ramasco, “Systematic comparison of trip distribution laws and models,” 06 2015.

[13] C. Robinson and B. N. Dilkina, "A machine learning approach to modeling human migration," *CoRR*, vol. abs/1711.05462, 2017.

[14] X. Lu and T. Zhao, "Research on time series data prediction based on clustering algorithm-a case study of yuebao," in *AIP Conference Proceedings*, vol. 1864, no. 1. AIP Publishing, 2017, p. 020152.

[15] X. Zhang, "Time series analysis and prediction by neural networks," *Optimization Methods and Software*, vol. 4, no. 2, pp. 151–170, 1994.

[16] B. Choi, "Applying machine learning methods for time series forecasting," 2009.

[17] N. Derby, "Time series forecasting methods," Statis Pro Data Analytics, Seattle, USA, Victoria SAS Users Group, 2008.

[18] H. Niska, T. Hiltunen, A. Karppinen, J. Ruuskanen, and M. Kolehmainen, "Evolving the neural network model for forecasting air pollution time series," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 2, pp. 159–167, 2004.

[19] S. Mahfoud and G. Mani, "Financial forecasting using genetic algorithms," *Applied artificial intelligence*, vol. 10, no. 6, pp. 543–566, 1996.

[20] G. Dorffner, "Neural networks for time series processing," *Neural Network World*, vol. 6, pp. 447–468, 1996.

[21] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[22] P. Palangpour, G. K. Venayagamoorthy, and K. Duffy, "Recurrent neural network based predictions of elephant migration in a south african game reserve," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 4084–4088.

[23] P. Cortez, M. Rocha, and J. Neves, "Genetic and evolutionary algorithms for time series forecasting," pp. 393–402, 06 2001.

[24] A. Freitas and E. Curry, *Big Data Curation*, J. M. Cavanillas, E. Curry, and W. Wahlster, Eds. Cham: Springer International Publishing, 2016.

[25] M. Aagaard, "How to deal with outliers in your data." [Online]. Available: https://conversionxl.com/blog/outliers/

[26] G. M. Kashif, A. Tirusew, K. Yasir, and M. Mac, "Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique," *Water Resources Research*, vol. 43, no. 7.

[27] K. H. Azme Khamis, Zuhaimy Ismail and A. T. Mohammed, "The effects of outliers data on neural network performance." *Journal of Applied Sciences*, vol. 5, pp. 1394–1398, 2005.

[28] C. A. Hosman, B. B. Hansen, and P. W. Holland, "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," 2009.

[29] FSNAU, "Somalia acute food insecurity situation overview, rural, urban and idp populations: February - june 2018, most likely scenario." [Online]. Available: http://www.fsnau.org/ipc/ipc-map

[30] D. Suleimenova, D. Bell, and D. Groen, "A generalized simulation development approach for predicting refugee destinations," in *Scientific Reports*, 2017.

Appendix A

Appendix

In this section are included all the models and statistics, for the readers interest, and for clarification of notions used in the main body of this report.

A.1 DATA

Date	Hiiraan_Belet_	Hiiraan_Bulo_	Shabelle_Dhexe_J	Juba_Dhexe_E	Gedo_Bardheere	Gedo_Luuq	Gedo_Dollo
4/1/2017	3.83	2.89	2.55	1.85	3.83	2.13	1.40
5/1/2017	5.05	4.20	4.99	3.65	5.05	2.88	
6/1/2017	5.20	2.58	4.82	3.52	5.20	1.78	
7/1/2017	4.37	1.56	2.53	2.45	4.37	1.24	
8/1/2017	4.91	2.49	3.58	2.74	4.91	1.86	
9/1/2017	5.76	4.23	5.05	4.79	5.76	2.55	
10/1/2017	8.59	4.76	4.96	5.82	6.76	3.86	
11/1/2017	4.55	4.03	4.73	6.24	7.00	4.92	
12/1/2017	3.35	2.47	1.86	2.85	4.56	0.20	0.62
1/1/2018	1.81	1.77	1.21	2.27	3.59	1.78	1.00
2/1/2018	1.71	1.14	0.60	1.87	3.72	1.44	1.00
3/1/2018	2.66	1.94	2.70	3.19	4.65	2.68	1.76
4/1/2018	2.17	1.59	1.53	2.39	4.36	2.26	1.88

Date	Awdal_Conflict	Bakool_Conflict	Banadir_Conflict	Bari_Conflict	Bay_Conflict	Galgaduu	Gedo_Co	Hiiraan_C
3/1/2017	2	3	66	11	13	11	8	6
4/1/2017	2	4	76	3	19	2	12	14
5/1/2017	0	5	42	11	15	5	12	19
6/1/2017	1	1	44	15	11	4	17	21
7/1/2017	1	6	50	9	9	6	17	27
8/1/2017	0	5	44	11	10	18	10	17
9/1/2017	3	4	57	13	12	2	21	12
10/1/2017	1	17	57	21	9	5	5	19
11/1/2017	3	6	41	17	8	4	2	9
12/1/2017	0	5	36	11	8	3	5	12
1/1/2018	3	4	55	20	22	5	8	23
2/1/2018	3	2	64	17	9	3	10	16
3/1/2018	1	2	61	37	9	5	14	17
4/1/2018	3	3	59	13	7	3	5	13

Figure A.1: Data on Conflict for 8 sample regions

A.2 GEA

DELAY -1 ALL Without Date, ALL, Until June -
Including Water Level, Rivers and Current location
1) Water Drum Price, 2) Local goat prices, 3)
Violent Incidents, 4) Fatalities, 5) Arrivals, 6)
Departures, 7) rainfall, 8) river discharge Training
50%/ and Validation 50% + Absolute error + YES
shuffle + maximum historical data 20%

Figure A.2: Experimental Set Up for GEA

MODEL	Best Fit	R^2	Max error
modelarrivals_BAminus1	0.202	0.96409686	12368.645
modelarrivals_BA1	0.282	0.87364666	23646.507
modelarrivals_BA2	0.285	0.91844285	15343.017
modelarrivals_BA3	0.311	0.88334732	22625.243
modelarrivals_BA4	0.274	0.91833948	16101.017
modelarrivals_BA5	0.285	0.90378999	19271.351
modelarrivals_BA6	0.309	0.90087281	21577.408
modelarrivals_BA7	0.275	0.89544755	26071.342
modelarrivals_BA8	0.318	0.86930941	24919.714
modelarrivals_BA9	0.312	0.8792886	21654.162
modelarrivals_BA10	0.245	0.93601843	11436.716
modelarrivals_BAminus2	0.217	0.94353182	11491.782
modelarrivals_BAJUN1	0.266	0.91521149	22638.666
modelarrivals_BAJUN2	0.265	0.93912173	14817.968
modelarrivals_BAJUN3	0.244	0.95018054	10548.816
modelarrivals_BAJUN4	0.244	0.95018052	10548.816
modelarrivals_BAJUN5	0.273	0.94062909	14572.362
modelarrivals_BAJUN6	0.297	0.90666278	21379.255
modelarrivals_BAJUN7	0.301	0.96605226	18325.5237
modelarrivals_BAJUN8	0.280	0.93636607	11495.035
modelarrivals_BAJUN9	0.274	0.93727837	14041.011
modelarrivals_BAJUN10	0.282	0.91090558	20926.657

Table A.1: Models and Fits

```
Banadir_CurrentRegion = f(delay(Awdal_Conflict, 1), delay(Bakool_Conflict, 1), delay(Banadir_Conflict, 1),
delay(Bari_Conflict, 1), delay(Bay_Conflict, 1), delay(Galgaduud_Conflict, 1), delay(Gedo_Conflict, 1),
delay(Hiiraan_Conflict, 1), delay(Jubbada_Dhexe_Conflict, 1), delay(Jubbada_Hoose_Conflict, 1),
delay(Mudug_Conflict, 1), delay(Nugaal_Conflict, 1), delay(Sanaag_Conflict, 1),
delay(Shabeellaha_Dhexe_Conflict, 1), delay(Shabeellaha_Hoose_Conflict, 1), delay(Sool_Conflict, 1),
delay(Togdheer_Conflict, 1), delay(Woqooyi_Galbeed_Conflict, 1), delay(Awdal_Fatalities, 1),
delay(Bakool_Fatalities, 1), delay(Banaadir_Fatalities, 1), delay(Bari_Fatalities, 1), delay(Bay_Fatalities, 1),
delay(Galguduud_Fatalities, 1), delay(Gedo_Fatalities, 1), delay(Hiiraan_Fatalities, 1),
delay(Jubbada_Dhexe_Fatalities, 1), delay(Jubbada_Hoose_Fatalities, 1), delay(Mudug_Fatalities, 1),
delay(Nugaal_Fatalities, 1), delay(Sanaag_Fatalities, 1), delay(Shabeellaha_Dhexe_Fatalities, 1),
delay(Shabeellaha_Hoose_Fatalities, 1), delay(Sool_Fatalities, 1), delay(Togdheer_Fatalities, 1),
delay(Woqooyi_Galbeed_Fatalities, 1), delay(Awdal_BeforeRegion, 1), delay(Bakool_BeforeRegion, 1),
delay(Banadir_BeforeRegion, 1), delay(Bari_BeforeRegion, 1), delay(Bay_BeforeRegion, 1),
delay(Galgaduud_BeforeRegion, 1), delay(Gedo_BeforeRegion, 1), delay(Hiiraan_BeforeRegion, 1),
delay(Jubbada_Dhexe_BeforeRegion, 1), delay(Jubbada_Hoose_BeforeRegion, 1), delay(Mudug_BeforeRegion,
1), delay(Nugaal_BeforeRegion, 1), delay(Sanaag_BeforeRegion, 1), delay(Shabeellaha_Dhexe_BeforeRegion, 1),
delay(Shabeellaha_Hoose_BeforeRegion, 1), delay(Sool_BeforeRegion, 1), delay(Togdheer_BeforeRegion, 1),
delay(Woqooyi_Galbeed_BeforeRegion, 1), delay(Awdal_FutureRegion, 1), delay(Bakool_FutureRegion, 1),
delay(Banadir_FutureRegion, 1), delay(Bari_FutureRegion, 1), delay(Bay_FutureRegion, 1),
delay(Galgaduud_FutureRegion, 1), delay(Gedo_FutureRegion, 1), delay(Hiiraan_FutureRegion, 1),
delay(Jubbada_Dhexe_FutureRegion, 1), delay(Jubbada_Hoose_FutureRegion, 1), delay(Mudug_FutureRegion,
1), delay(Nugaal_FutureRegion, 1), delay(Sanaag_FutureRegion, 1), delay(Shabeallaha_Dhexe_FutureRegion, 1),
delay(Shabeellaha_Hoose_FutureRegion, 1), delay(Sool_FutureRegion, 1), delay(Togdheer_FutureRegion, 1),
delay(Woqooyi_Galbeed_FutureRegion, 1), delay(Juba_River_discharge, 1), delay(Shabelle_River_discharge, 1),
delay(Hiiraan_Belet_WeyneStation_Shabelle_River, 1), delay(Hiiraan_Bulo_Burti_StationShabelle_River, 1),
delay(Shabelle_Dhexe_JowharStation_Shabelle_River, 1), delay(Juba_Dhexe_BualleStation_Juba_River, 1),
delay(Gedo_BardheereStation_Juba_River, 1), delay(Gedo_LuuqStation_Juba_River, 1),
delay(Gedo_DollowStation_Juba_River, 1), delay(Awdal_WaterDrumPrice, 1), delay(Bakool_WaterDrumPrice, 1),
delay(Banadir_WaterDrumPrice, 1), delay(Bari_WaterDrumPrice, 1), delay(Bay_WaterDrumPrice, 1),
delay(Galgaduud_WaterDrumPrice, 1), delay(Gedo_WaterDrumPrice, 1), delay(Hiiraan_WaterDrumPrice, 1),
delay(Jubbada_Dhexe_WaterDrumPrice, 1), delay(Jubbada_Hoose_WaterDrumPrice, 1),
delay(Mudug_WaterDrumPrice, 1), delay(Nugaal_WaterDrumPrice, 1), delay(Sanaag_WaterDrumPrice, 1),
delay(Shabeallaha_Dhexe_WaterDrumPrice, 1), delay(Shabeellaha_Hoose_WaterDrumPrice, 1),
delay(Sool_WaterDrumPrice, 1), delay(Togdheer_WaterDrumPrice, 1), delay(Woqooyi_Galbeed_WaterDrumPrice,
1), delay(Awdal_goatprice, 1), delay(Bakool_goatprice, 1), delay(Banadir_goatprice, 1), delay(Bari_goatprice, 1),
delay(Bay_goatprice, 1), delay(Galgaduud_goatprice, 1), delay(Gedo_goatprice, 1), delay(Hiiraan_goatprice, 1),
delay(Jubbada_Dhexe_goatprice, 1), delay(Jubbada_Hoose_goatprice, 1), delay(Mudug_goatprice, 1),
delay(Nugaal_goatprice, 1), delay(Sanaag_goatprice, 1), delay(Shabeallaha_Dhexe_goatprice, 1),
delay(Shabeellaha_Hoose_goatprice, 1), delay(Sool_goatprice, 1), delay(Togdheer_goatprice, 1),
delay(Woqooyi_Galbeed_goatprice, 1), delay(Awdal_rain, 1), delay(Bakool_rain, 1), delay(Banaadir_rain, 1),
delay(Bari_rain, 1), delay(Bay_rain, 1), delay(Gedo_rain, 1), delay(Hiiraan_rain, 1), delay(Jubbada_Dhexe_rain, 1),
delay(Jubbada_Hoose_rain, 1), delay(Mudug_rain, 1), delay(Nugaal_rain, 1), delay(Sanaag_rain, 1),
delay(Shabeellaha_Dhexe_rain, 1), delay(Shabeellaha_Hoose_rain, 1), delay(Sool_rain, 1), delay(Togdheer_rain,
1), delay(Woqooyi_Galbeed_rain, 1), delay(Awdal_CurrentRegion, 1), delay(Bakool_CurrentRegion, 1),
delay(Bari_CurrentRegion, 1), delay(Galgaduud_CurrentRegion, 1), delay(Bay_CurrentRegion, 1),
delay(Hiiraan_CurrentRegion, 1), delay(Jubbada_Dhexe_CurrentRegion, 1), delay(Jubbada_Hoose_CurrentRegion,
1), delay(Mudug_CurrentRegion, 1), delay(Nugaal_CurrentRegion, 1), delay(Sanaag_CurrentRegion, 1),
delay(Shabeellaha_Dhexe_CurrentRegion, 1), delay(Shabeellaha_Hoose_CurrentRegion, 1),
delay(Sool_CurrentRegion, 1), delay(Togdheer_CurrentRegion, 1), delay(Woqooyi_Galbeed_CurrentRegion, 1),
delay(Gedo_CurrentRegion, 1))
```

Figure A.3: Function to Run on GEA

Banadir_CurrentRegion = 0.0932582045922656*delay(Bari_Fatalities, 1)*delay(Mudug_BeforeRegion, 1) + sqrt(delay(Woqooyi_Galbeed_FutureRegion, 1))*sma(delay(Nugaal_CurrentRegion, 1), 3) + delay(Gedo_CurrentRegion, 1)*floor(0.00207485985528903*delay(Bari_CurrentRegion, 1)) + delay(Gedo_Fatalities, 1)*mma(delay(Mudug_Fatalities, 1), 7)*smm(delay(Bakool_rain, 1), 8) + wma(delay(Hiiraan_BeforeRegion, 1), 12) + max(delay(Gedo_CurrentRegion, 1), 0.442050688064606*wma(delay(Bakool_WaterDrumPrice, 1), 4)) - 9060.67992987461	modelarrivals_BA1
Banadir_CurrentRegion = 0.820777499965034*delay(Bakool_WaterDrumPrice, 1) + delay(Gedo_BeforeRegion, 1)*delay(Sanaag_Fatalities, 3) + 0.000385779996089999*delay(Mudug_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1) + delay(Gedo_BeforeRegion, 1)*sin(delay(Banadir_WaterDrumPrice, 11)) + 0.379038413366182*delay(Gedo_BeforeRegion, 1)*delay(Togdheer_Fatalities, 3) + delay(Sanaag_CurrentRegion, 1)*cos(0.820777499965034*delay(Bakool_WaterDrumPrice, 1)) + delay(Jubbada_Dhexe_CurrentRegion, 15) - 0.0144876933172979*delay(Jubbada_Hoose_goatprice, 1)	modelarrivals_BA2
Banadir_CurrentRegion = 0.236728395075358*delay(Hiiraan_BeforeRegion, 1) + 0.00370133425867487*delay(Hiiraan_BeforeRegion, 1)*delay(Mudug_FutureRegion, 1) + 0.0116396096993031*delay(Galgaduud_FutureRegion, 1)*mma(delay(Mudug_BeforeRegion, 1), 2) + delay(Shabeellaha_Dhexe_BeforeRegion, 1) + delay(Mudug_FutureRegion, 14) + wma(delay(Bakool_WaterDrumPrice, 1), 11) - 19896.2179632592 - 1.9071651571991*delay(Gedo_CurrentRegion, 1)*round(-0.00103838434135449*delay(Bari_CurrentRegion, 1))	modelarrivals_BA3
Banadir_CurrentRegion = delay(Sanaag_Conflict, 3)*if(delay(Nugaal_Conflict, 1), delay(Juba_River_discharge, 11), delay(Gedo_CurrentRegion, 1)) + max(0.694684361700432*delay(Bakool_WaterDrumPrice, 1) + delay(Togdheer_CurrentRegion, 1)*delay(Gedo_DollowStation_Juba_River, 2) + delay(Hiiraan_Conflict, 9)*delay(Togdheer_CurrentRegion, 9) + delay(Bay_BeforeRegion, 1) + sma(delay(Mudug_BeforeRegion, 1), 2), delay(Galgaduud_WaterDrumPrice, 7)) - 0.0162014099906871*mma(delay(Jubbada_Hoose_goatprice, 1), 15) - 5.90971441685906*mma(delay(Togdheer_BeforeRegion, 1), 2)	modelarrivals_BA4
Banadir_CurrentRegion = 0.704821320693777*delay(Bakool_WaterDrumPrice, 1) + 0.000402906694472956*delay(Mudug_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1) + delay(Jubbada_Hoose_FutureRegion, 15)*mma(delay(Togdheer_Fatalities, 1), 10) + delay(Sanaag_CurrentRegion, 1)*sin(0.704821320693777*delay(Bakool_WaterDrumPrice, 1)) + delay(Jubbada_Dhexe_CurrentRegion, 15) + min(26734.1634362641, delay(Sanaag_Fatalities, 3)*round(1.92235530765 21*delay(Gedo_BeforeRegion, 1))) - 0.0122243904473122*delay(Jubbada_Hoose_goatprice, 1)	modelarrivals_BA5

Figure A.4: Models of GEA for September 1 to 5

Banadir_CurrentRegion = 0.656699238779318*delay(Bay_BeforeRegion, 1) + 0.656699238779318*delay(Bakool_WaterDrumPrice, 1) + 0.494683614858048*delay(Mudug_BeforeRegion, 1) + 0.751616373213706*delay(Gedo_CurrentRegion, 1)*delay(Sanaag_Fatalities, 3) + wma(delay(Hiiraan_Bulo_Burti_StationShabelle_River, 1), 14)*min(delay(Bay_BeforeRegion, 1), delay(Mudug_BeforeRegion, 1)) + delay(Galgaduud_FutureRegion, 1) + smm(delay(Galgaduud_BeforeRegion, 1), 4) - 3.83040093927554*delay(Togdheer_BeforeRegion, 1) - 0.011783920876017*sma(delay(Gedo_goatprice, 1), 3)	modelarrivals_BA6
Banadir_CurrentRegion = 5.45258316814397*delay(Mudug_BeforeRegion, 1) + 0.634530258083762*delay(Bakool_WaterDrumPrice, 1) + delay(Gedo_BeforeRegion, 1)*sma(delay(Jubbada_Dhexe_Conflict, 1), 15) + delay(Bari_FutureRegion, 1) + delay(Togdheer_WaterDrumPrice, 4) - delay(Mudug_BeforeRegion, 1)*min(delay(Nugaal_Conflict, 1), min(wma(delay(Gedo_rain, 1), 11), min(0.634530258083762*delay(Bakool_WaterDrumPrice, 1) - 9764.89637118794, delay(Bari_FutureRegion, 1)))) - 0.0138771809011655*delay(Jubbada_Hoose_goatprice, 1) - 5.38123507880934*delay(Gedo_FutureRegion, 10)	modelarrivals_BA7
Banadir_CurrentRegion = delay(Bari_FutureRegion, 1)*delay(Shabeellaha_Hoose_rain, 1) + 0.00197058875505935*delay(Bari_CurrentRegion, 1)*delay(Gedo_CurrentRegion, 1) + 8.48543352535101e-6*delay(Bakool_WaterDrumPrice, 1)^2 + 0.0120431601069705*delay(Mudug_BeforeRegion, 1)*mma(delay(Galgaduud_FutureRegion, 1), 4) + 0.00111435882011818*delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_BeforeRegion, 1), 5) + delay(Bakool_CurrentRegion, 6) + delay(Mudug_FutureRegion, 14) - sma(delay(Jubbada_Hoose_BeforeRegion, 1), 7)	modelarrivals_BA8
Banadir_CurrentRegion = 0.371143649329236*delay(Bay_WaterDrumPrice, 1) + 0.0092141903043108*delay(Mudug_BeforeRegion, 1)*delay(Galgaduud_FutureRegion, 1) + 0.000977607114690172*delay(Bari_CurrentRegion, 1)*delay(Gedo_CurrentRegion, 1) + smm(delay(Shabeellaha_Dhexe_Fatalities, 1), 8)*smm(delay(Jubbada_Dhexe_CurrentRegion, 1), 13) + delay(Hiiraan_BeforeRegion, 1)*wma(delay(Gedo_Conflict, 1), 3)*atan2(0.000977607114690172*delay(Bari_CurrentRegion, 1)*delay(Gedo_CurrentRegion, 1), delay(Bay_WaterDrumPrice, 1)) + delay(Nugaal_FutureRegion, 3) + sma(delay(Gedo_CurrentRegion, 1), 9) - 0.0092141903043108*delay(Jubbada_Hoose_goatprice, 1)	modelarrivals_BA9
Banadir_CurrentRegion = 0.000400206821442192*delay(Mudug_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1) + delay(Sanaag_Fatalities, 3)*delay(Nugaal_CurrentRegion, 3)*wma(delay(Bari_rain, 1), 11) + delay(Nugaal_FutureRegion, 3) + delay(Nugaal_FutureRegion, 4) + max(6329.56723946874, delay(Sanaag_CurrentRegion, 1)) + max(0.490846430861034*delay(Bakool_WaterDrumPrice, 1) + 130.37533697164*wma(delay(Shabeellaha_Dhexe_Fatalities, 1), 2), delay(Mudug_BeforeRegion, 1)) - 0.0152068781052589*sma(delay(Gedo_goatprice, 1), 4)	modelarrivals_BA10

Figure A.5: Models of GEA for September 5 to 10

Banadir_CurrentRegion = 0.0868703284495857*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + 0.0819722756534686*delay(Bari_Conflict, 1)*mma(delay(Banadir_BeforeRegion, 1), 3) + 3.20101567484822*sma(delay(Togdheer_Fatalities, 1), 3)*wma(delay(Gedo_FutureRegion, 1), 5) + 4.89025816928898*delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_BeforeRegion, 1), 7)/(0.0868703284495857*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + 0.0819722756534686*delay(Bari_Conflict, 1)*mma(delay(Banadir_BeforeRegion, 1), 3) + 3.20101567484822*sma(delay(Togdheer_Fatalities, 1), 3)*wma(delay(Gedo_FutureRegion, 1), 5)) + max(0.586149473609031*delay(Hiiraan_WaterDrumPrice, 1), delay(Mudug_BeforeRegion, 1)) - 8523.16781354588	modelarrivals_BAJUN1
Banadir_CurrentRegion = 0.0732602367252228*delay(Bari_Conflict, 1)*delay(Banadir_BeforeRegion, 1) + 0.0732602367252228*delay(Bari_Conflict, 1)*delay(Mudug_BeforeRegion, 1) + 0.0732602367252228*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + 8.54019598974934e-6*delay(Bakool_WaterDrumPrice, 1)*delay(Hiiraan_WaterDrumPrice, 1) + delay(Shabeellaha_Dhexe_BeforeRegion, 6) + max(0.177793684635044*delay(Hiiraan_BeforeRegion, 1), 1.18213777130181*delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_rain, 1), 14)) - delay(Mudug_CurrentRegion, 17)	modelarrivals_BAJUN2
Banadir_CurrentRegion = 0.00644970554752389*delay(Hiiraan_BeforeRegion, 1)*delay(Nugaal_CurrentRegion, 1) + 0.000232134354975325*delay(Mudug_BeforeRegion, 1)*delay(Bay_WaterDrumPrice, 1) + delay(Sool_CurrentRegion, 4)*delay(Awdal_Conflict, 6) + delay(Shabeellaha_Dhexe_BeforeRegion, 6)*delay(Sanaag_Fatalities, 9) + delay(Nugaal_FutureRegion, 1) + delay(Bay_BeforeRegion, 8) + sma(delay(Banadir_BeforeRegion, 1), 3) + exp(delay(Awdal_Conflict, 6)) - delay(Hiiraan_BeforeRegion, 1) - delay(Nugaal_CurrentRegion, 1) - 4.81297710245971*delay(Mudug_BeforeRegion, 1)	modelarrivals_BAJUN3
Banadir_CurrentRegion = 0.00644968999560936*delay(Hiiraan_BeforeRegion, 1)*delay(Nugaal_CurrentRegion, 1) + 0.000232134354975325*delay(Mudug_BeforeRegion, 1)*delay(Bay_WaterDrumPrice, 1) + delay(Sool_CurrentRegion, 4)*delay(Awdal_Conflict, 6) + delay(Shabeellaha_Dhexe_BeforeRegion, 6)*delay(Sanaag_Fatalities, 9) + delay(Nugaal_FutureRegion, 1) + delay(Bay_BeforeRegion, 8) + sma(delay(Banadir_BeforeRegion, 1), 3) + exp(delay(Awdal_Conflict, 6)) - delay(Hiiraan_BeforeRegion, 1) - delay(Nugaal_CurrentRegion, 1) - 4.81297710245971*delay(Mudug_BeforeRegion, 1)	modelarrivals_BAJUN4
Banadir_CurrentRegion = max(delay(Mudug_BeforeRegion, 1), 0.464844464717392*delay(Hiiraan_WaterDrumPrice, 1) + 0.156038424901233*delay(Hiiraan_BeforeRegion, 1) + delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_rain, 1), 14) + 0.0936984664666861*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + 0.0767994494620685*delay(Bari_Conflict, 1)*delay(Banadir_BeforeRegion, 1) - delay(Jubbada_Dhexe_FutureRegion, 1)) - delay(Nugaal_WaterDrumPrice, 7) - delay(Awdal_BeforeRegion, 1)*delay(Woqooyi_Galbeed_rain, 1)	modelarrivals_BAJUN5

Figure A.6: Models of GEA for June 1 to 5

Banadir_CurrentRegion = 0.500464772441319*delay(Hiiraan_WaterDrumPrice, 1) + 0.0773220425411127*delay(Bari_Conflict, 1)*delay(Banadir_BeforeRegion, 1) + sma(delay(Togdheer_Fatalities, 1), 2)*mma(delay(Gedo_FutureRegion, 1), 5) + 0.0134917409605303*delay(Hiiraan_BeforeRegion, 1)*delay(Nugaal_CurrentRegion, 3) + 0.00795798570650005*delay(Bari_Conflict, 1)*delay(Mudug_Fatalities, 1)*mma(delay(Bay_BeforeRegion, 1), 2) + mma(delay(Gedo_FutureRegion, 1), 5) - 5721.92313728203 - 2.44915092407495*delay(Hiiraan_BeforeRegion, 1)	modelarrivals_BAJUN6
Banadir_CurrentRegion = delay(Shabeellaha_Dhexe_BeforeRegion, 15) + mma(delay(Mudug_BeforeRegion, 1), 6) + smm(delay(Banadir_BeforeRegion, 1), 15) + max(0.0745988338790564*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + 0.0679902969578057*delay(Bari_Conflict, 1)*delay(Banadir_BeforeRegion, 1) - delay(Nugaal_Fatalities, 1)*sin(0.0963804584498508*delay(Mudug_Fatalities, 1))*if(cos(delay(Bari_Conflict, 1)), 0.0745988338790564*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1), delay(Bay_BeforeRegion, 1)), 0.00110933805415979*mma(delay(Hiiraan_CurrentRegion, 1), 3)*mma(delay(Gedo_BeforeRegion, 1), 9))	modelarrivals_BAJUN7
Banadir_CurrentRegion = max(0.525731462646251*delay(Hiiraan_WaterDrumPrice, 1) + 0.164551229335035*delay(Hiiraan_BeforeRegion, 1) + delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_rain, 1), 14) + 0.0772156733729328*delay(Bari_Conflict, 1)*delay(Banadir_BeforeRegion, 1) + delay(Nugaal_CurrentRegion, 1) + smm(delay(Nugaal_CurrentRegion, 1), 4) + min(delay(Hiiraan_BeforeRegion, 1)*smm(delay(Nugaal_CurrentRegion, 1), 4), 0.0938139964853952*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1)) - 6252.41542789858, delay(Gedo_CurrentRegion, 8))	modelarrivals_BAJUN8
Banadir_CurrentRegion = 11723.7642528514 + 0.00790448542053325*delay(Mudug_BeforeRegion, 1)*delay(Galgaduud_FutureRegion, 1) + 0.0680140693765127*delay(Bari_Conflict, 1)*mma(delay(Banadir_BeforeRegion, 1), 3) + 0.0016221206705837*delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_BeforeRegion, 1), 5) + 0.0813222729835992*delay(Bay_BeforeRegion, 1)*min(0.0016221206705837*delay(Hiiraan_BeforeRegion, 1)*smm(delay(Gedo_BeforeRegion, 1), 5), mma(delay(Mudug_Fatalities, 1), 3)) + delay(Nugaal_FutureRegion, 8) - 0.00920377145047265*mma(delay(Gedo_goatprice, 1), 7)	modelarrivals_BAJUN9
Banadir_CurrentRegion = delay(Bakool_FutureRegion, 1)*delay(Gedo_DollowStation_Juba_River, 3) + 0.0695347098777455*delay(Mudug_Fatalities, 1)*delay(Bay_BeforeRegion, 1) + delay(Bay_BeforeRegion, 8)*delay(Hiiraan_Belet_WeyneStation_Shabelle_River, 9) + delay(Shabeellaha_Dhexe_BeforeRegion, 6)*atan2(delay(Nugaal_FutureRegion, 1), delay(Shabeellaha_Dhexe_rain, 1)) + delay(Bari_CurrentRegion, 3) + sma(delay(Banadir_BeforeRegion, 1), 3) + wma(delay(Mudug_BeforeRegion, 1), 2) + max(delay(Nugaal_FutureRegion, 1), 0.33261235258023*mma(delay(Hiiraan_BeforeRegion, 1), 3)) - smm(delay(Galgaduud_CurrentRegion, 1), 8)	modelarrivals_BAJUN10

Figure A.7: Models of GEA for June 5 to 10

Banadir_CurrentRegion = max(max(sma(delay(Shabeellaha_Hoose_CurrentRegion, 1), 4), delay(Awdal_CurrentRegion, 6)*sma(delay(Galgaduud_Conflict, 1), 7) + 0.000313298891221263*delay(Gedo_CurrentRegion, 1)*mma(delay(Mudug_BeforeRegion, 1), 4) + delay(Bay_BeforeRegion, 1) + delay(Bakool_WaterDrumPrice, 1) + delay(Bakool_CurrentRegion, 1) + delay(Banadir_FutureRegion, 3) + delay(Woqooyi_Galbeed_CurrentRegion, 11) - 0.0258563581198834*sma(delay(Jubbada_Hoose_goatprice, 1), 2)), max(mma(delay(Mudug_BeforeRegion, 1), 4), delay(Awdal_WaterDrumPrice, 1))) - 5.09700429342906*delay(Togdheer_BeforeRegion, 1)	BA_JAN1
Banadir_CurrentRegion = 0.758431461970115*delay(Bakool_WaterDrumPrice, 1) + 0.114269094533391*delay(Hiiraan_CurrentRegion, 1) + 0.0318884657497103*delay(Bari_BeforeRegion, 1)*delay(Bari_FutureRegion, 1) + 0.00164281143615443*delay(Gedo_BeforeRegion, 1)*delay(Jubbada_Hoose_CurrentRegion, 1) + delay(Sanaag_BeforeRegion, 1)/cos(delay(Banaadir_Fatalities , 1)) + delay(Nugaal_FutureRegion, 3) + mma(delay(Mudug_BeforeRegion, 1), 10) - 0.0132551605794407*delay(Jubbada_Hoose_goatprice, 1)	BA_JAN2
Banadir_CurrentRegion = 0.42741960289751*delay(Bakool_WaterDrumPrice, 1) + delay(Togdheer_BeforeRegion, 4)*smm(delay(Bari_rain, 1), 12) + delay(Bay_BeforeRegion, 1) + delay(Gedo_CurrentRegion, 1) + delay(Woqooyi_Galbeed_CurrentRegion, 11) + delay(Sool_BeforeRegion, 12) + delay(Sool_CurrentRegion, 13) + delay(Mudug_BeforeRegion, 15) + delay(Shabeellaha_Dhexe_BeforeRegion, 15) + mma(delay(Mudug_BeforeRegion, 1), 2) - 0.00857977637 16422*delay(Bakool_goatprice, 1) - 4.1555144852 1433*delay(Togdheer_BeforeRegion, 1)	BA_JAN3
Banadir_CurrentRegion = 0.723659599277594*mma(delay(Bakool_WaterDrumPrice, 1), 3) + delay(Shabeellaha_Hoose_rain, 2)*mma(delay(Togdheer_FutureRegion, 1), 5) + delay(Togdheer_CurrentRegion, 10)*sma(delay(Awdal_Conflict, 1), 16) + 0.723659599277594*delay(Gedo_CurrentRegion, 1)*delay(Sanaag_Fatalities, 3) + 4.19259262553008*delay(Bay_BeforeRegion, 1)*atan2(delay(Mudug_CurrentRegion, 1), mma(delay(Bay_CurrentRegion, 1), 3)) - delay(Woqooyi_Galbeed_Fatalities , 1)*delay(Togdheer_BeforeRegion, 1) - 0.0129056930227057*mma(delay(Jubbada_Hoose_goatprice, 1), 17)	BA_JAN4
Banadir_CurrentRegion = max(delay(Galgaduud_FutureRegion, 1)*delay(Woqooyi_Galbeed_Conflict, 4) + 0.0226127865483693*delay(Bari_BeforeRegion, 1)*delay(Bari_FutureRegion, 1) + delay(Jubbada_Hoose_BeforeRegion, 1)*round(0.00328503940216911*delay(Gedo_BeforeRegion, 1)) + delay(Woqooyi_Galbeed_CurrentRegion, 3) + mma(delay(Bakool_WaterDrumPrice, 1), 4) + mma(delay(Mudug_BeforeRegion, 1), 8) + mod(delay(Banadir_WaterDrumPrice, 1), smm(delay(Mudug_WaterDrumPrice, 1), 2)) - 0.0259449432621717*delay(Jubbada_Hoose_goatprice, 1), delay(Nugaal_WaterDrumPrice, 1))	BA_JAN5

Figure A.8: Models of GEA for January 1 to 5

Banadir_CurrentRegion = max(sma(delay(Shabeellaha_Hoose_CurrentRegion, 1), 4), 0.694309676539598*delay(Bakool_WaterDrumPrice, 1) + 4.40959184539803*delay(Awdal_CurrentRegion, 6) + delay(Mudug_Conflict, 1)*delay(Bari_BeforeRegion, 2) + delay(Hiiraan_Fatalities, 1)*delay(Togdheer_rain, 2) + delay(Bay_BeforeRegion, 1) + smm(delay(Mudug_BeforeRegion, 1), 2) + mod(sin(0.694309676539598*delay(Bakool_WaterDrumPrice, 1) + 4.40959184539803*delay(Awdal_CurrentRegion, 6) + delay(Bay_BeforeRegion, 1)), delay(Bay_BeforeRegion, 1)) - 0.0132818010194002*delay(Jubbada_Hoose_goatprice, 1)) - 5.30309570083848*delay(Togdheer_BeforeRegion, 1)	BA_JAN6
Banadir_CurrentRegion = 0.383773319042241*delay(Mudug_BeforeRegion, 1) + 0.000254407861552372*delay(Mudug_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1) + 0.617280788506247*delay(Gedo_CurrentRegion, 1)*delay(Sanaag_Fatalities, 3) + 2.52131879555939e-9*delay(Bakool_WaterDrumPrice, 1)^2*delay(Togdheer_WaterDrumPrice, 1) + delay(Bay_BeforeRegion, 1) + delay(Shabeallaha_Dhexe_WaterDrumPrice, 5) - 0.0165203257108142*delay(Jubbada_Hoose_goatprice, 1) - 5.29797410507584*delay(Togdheer_BeforeRegion, 1)	BA_JAN7
Banadir_CurrentRegion = 1.38222885637822*delay(Bay_Fatalities, 1)*delay(Shabeellaha_Dhexe_Fatalities, 1) + 0.000625349024349665*delay(Bari_BeforeRegion, 1)*delay(Sool_BeforeRegion, 1) + mma(delay(Mudug_BeforeRegion, 1), 9) + mma(delay(Bakool_WaterDrumPrice, 1), 11) + max(0.0291127874227999*delay(Bari_BeforeRegion, 1)*delay(Bari_FutureRegion, 1), 0.0031484011693576*delay(Gedo_BeforeRegion, 1)*delay(Jubbada_Hoose_BeforeRegion, 1)) - delay(Shabeellaha_Dhexe_BeforeRegion, 5) - 0.0187859943849533*delay(Jubbada_Hoose_goatprice, 1)	BA_JAN8
Banadir_CurrentRegion = 0.386917311805733*delay(Bakool_WaterDrumPrice, 1) + 0.000164659933885506*delay(Bay_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1) + 1.95859170065185*if(greater_or_equal(delay(Bari_rain, 1), less(0.386917311805733, 0.000164659933885506*delay(Bay_BeforeRegion, 1)*delay(Gedo_CurrentRegion, 1))), sma(delay(Mudug_BeforeRegion, 1), 11), delay(Shabeellaha_Dhexe_BeforeRegion, 6)) + delay(Gedo_CurrentRegion, 6) + delay(Bay_BeforeRegion, 8) + delay(Sool_CurrentRegion, 13) + mma(delay(Mudug_BeforeRegion, 1), 3) + mma(delay(Shabeellaha_Dhexe_BeforeRegion, 1), 5) - 0.00925377719973352*delay(Bakool_goatprice, 1)	BA_JAN9
Banadir_CurrentRegion = 0.43034473721445*delay(Bay_BeforeRegion, 1) + 3.14493060846319*min(delay(Bay_CurrentRegion, 1), delay(Mudug_BeforeRegion, 1)) + 0.702705597861009*delay(Gedo_CurrentRegion, 1)*delay(Sanaag_Fatalities, 3) + delay(Togdheer_CurrentRegion, 9) + mma(delay(Bakool_CurrentRegion, 1), 3) + max(delay(Mudug_BeforeRegion, 1), 0.719087384261325*delay(Bakool_WaterDrumPrice, 1)) - 0.0120073174431092*delay(Jubbada_Hoose_goatprice, 1) - 3.26356198195501*delay(Togdheer_BeforeRegion, 1)	BA_JAN10

Figure A.9: Models of GEA for January 5 to 10

SEP TRAIN	BA1	BA2	BA3
MAE	2657.78	2699.26	2710.59
MSE	24640300.00	16890600.00	20724600.00
R2	0.87	0.91	0.89
Correlation Coefficient	0.94	0.96	0.95
Rank Correlation	0.86	0.81	0.83
Max Error	23646.50	15343.00	21563.60
Log Error	14.07	13.82	13.56
Median Error	1291.23	1159.44	1596.25
Inter-quartile Absolute Error	1386.34	1611.09	1774.18
Signed Difference Error	571.92	21.73	263.49
HCE	0.41	0.34	0.37
AIC	1441.51	1405.12	1406.41

Table A.2: TRAINING SET METRICS SEPTEMBER MODELS part1

Table A.3: My caption

SEP TRAIN	BA4	BA5	BA6	BA7
MAE	2524.86	2666.88	2846.88	2639.60
MSE	15771800.00	19418600.00	19461300.00	21651000.00
R2	0.92	0.90	0.90	0.89
Correlation Coefficient	0.96	0.95	0.95	0.94
Rank Correlation	0.84	0.80	0.79	0.86
Max Error	16101.00	19271.40	21577.40	26071.30
Log Error	13.90	13.38	13.71	13.61
Median Error	1291.26	1488.66	1751.68	1507.13
Inter-quartile Absolute Error	1465.93	1567.51	2048.17	1615.62
Signed Difference Error	-210.77	-178.82	699.50	295.40
HCE	0.31	0.36	0.37	0.38
AIC	1399.91	1431.52	1440.25	1487.93

Table A.4: TRAINING SET METRICS SEPTEMBER MODELS part2

SEP TRAIN	BA8	BA9	BA10
MAE	2890.69	2841.23	2305.76
MSE	25163400.00	23290900.00	12511000.00
R2	0.87	0.88	0.94
Correlation Coefficient	0.93	0.94	0.97
Rank Correlation	0.86	0.90	0.86
Max Error	24919.70	21654.20	11436.70
Log Error	14.04	13.91	13.14
Median Error	1781.51	1584.54	996.61
Inter-quartile Absolute Error	1690.19	1613.49	1434.43
Signed Difference Error	284.80	799.06	218.90
HCE	0.43	0.41	0.27
AIC	1414.54	1462.95	1382.31

Table A.5: TRAINING SET METRICS SEPTEMBER MODELS part3

JUNE TRAIN	BAJUN1	BAJUN2	BAJUN3
MAE	2448.29	2708.26	2130.53
MSE	19393200.00	20342100.00	10506100.00
R2	0.90	0.89	0.95
Correlation Coefficient	0.95	0.95	0.97
Rank Correlation	0.87	0.79	0.84
Max Error	22638.70	19601.60	10548.80
Log Error	13.55	13.72	13.36
Median Error	1262.30	1385.63	1421.66
Inter-quartile Absolute Error	1244.22	1656.03	1383.02
Signed Difference Error	-17.50	216.66	626.32
HCE	0.34	0.38	0.25
AIC	1697.32	1374.46	1303.76

Table A.6: TRAINING SET METRICS JUNE MODELS part1

JUNE TRAIN	BAJUN4	BAJUN5	BAJUN6	BAJUN7
MAE	2130.53	2689.51	2692.37	2320.30
MSE	10506100.00	19734400.00	21045100.00	13294300.00
R2	0.95	0.90	0.89	0.93
Correlation Coefficient	0.97	0.95	0.95	0.97
Rank Correlation	0.84	0.81	0.85	0.85
Max Error	10548.80	20383.20	21379.30	16396.90
Log Error	13.36	13.70	13.54	13.68
Median Error	1421.65	1669.24	1546.35	1451.63
Inter-quartile Absolute Error	1383.03	1669.20	1532.32	1564.27
Signed Difference Error	626.33	651.88	792.40	522.79
HCE	0.25	0.37	0.38	0.28
AIC	1303.76	1356.88	1394.12	1413.89

Table A.7: TRAINING SET METRICS JUNE MODELS part2

JUNE TRAIN	BAJUN8	BAJUN9	BAJUN10
MAE	2603.44	2652.13	2539.08
MSE	15955100.00	19628900.00	20368000.00
R2	0.92	0.90	0.89
Correlation Coefficient	0.96	0.95	0.95
Rank Correlation	0.84	0.86	0.89
Max Error	13182.50	18288.40	20926.70
Log Error	13.78	13.43	13.69
Median Error	1666.19	1142.17	1149.67
Inter-quartile Absolute Error	1660.24	1519.12	1152.35
Signed Difference Error	296.31	465.74	-433.39
HCE	0.33	0.36	0.35
AIC	1365.07	1418.95	1352.09

Table A.8: TRAINING SET METRICS JUNE MODELS part3

BA_10 Variables	Sensitivity	% Positive	PM	% Negative	NM
Sanaag_Fatalities	0.31	100%	0.31	0%	0
Bari_rain	0.28	100%	0.28	0%	0
Bakool_WaterDrumPrice	0.26	100%	0.26	0%	0
Gedo_goatprice	0.22	0%	0.00	100%	0.2
Nugaal_CurrentRegion	0.22	100%	0.22	0%	0
Mudug_BeforeRegion	0.18	100%	0.18	0%	0
Shabeellaha_Dhexe_Fatalities	0.14	100%	0.14	0%	0
Gedo_CurrentRegion	0.10	100%	0.10	0%	0
Nugaal_FutureRegion	0.05	100%	0.05	0%	0
Sanaag_CurrentRegion	0.01	100%	0.01	0%	0
BA_9 Variables	Sensitivity	% Positive	PM	% Negative	NM
Shabeellaha_Dhexe_Fatalities	0.64	100%	0.64	0%	0
Jubbada_Dhexe_CurrentRegion	0.48	100%	0.48	0%	0
Mudug_BeforeRegion	0.44	100%	0.44	0%	0
Gedo_CurrentRegion	0.44	100%	0.44	0%	0
Bari_CurrentRegion	0.43	100%	0.43	0%	0
Galgaduud_FutureRegion	0.39	100%	0.39	0%	0
Hiiraan_BeforeRegion	0.15	100%	0.15	0%	0
Jubbada_Hoose_goatprice	0.15	0%	0.00	100%	0.15
Bay_WaterDrumPrice	0.12	100%	0.12	0%	0
Gedo_Conflict	0.05	100%	0.05	0%	0
Nugaal_FutureRegion	0.03	100%	0.03	0%	0
BA_8 Variables	Sensitivity	% Positive	PM	% Negative	NM
Mudug_BeforeRegion	0.55	100%	0.55	0%	0
Galgaduud_FutureRegion	0.48	100%	0.48	0%	0
Bari_CurrentRegion	0.38	100%	0.38	0%	0
Mudug_FutureRegion	0.30	100%	0.30	0%	0
Gedo_CurrentRegion	0.29	100%	0.29	0%	0
Bakool_WaterDrumPrice	0.25	100%	0.25	0%	0
Bari_FutureRegion	0.22	100%	0.22	0%	0
Gedo_BeforeRegion	0.19	100%	0.19	0%	0
Hiiraan_BeforeRegion	0.17	100%	0.17	0%	0
Shabeellaha_Hoose_rain	0.15	100%	0.15	0%	0
Bakool_CurrentRegion	0.13	100%	0.13	0%	0
Jubbada_Hoose_BeforeRegion	0.12	0%	0.00	100%	0.12

Table A.9: Sensitivity of Models for September in descending order 10-8

BA_7 Variables	Sensitivity	% Positive	PM	% Negative	NM
Bakool_WaterDrumPrice	5.90	97%	0.38	3%	182.56
Mudug_BeforeRegion	2.10	100%	2.10	0%	0.00
Bari_FutureRegion	1.24	85%	0.02	15%	7.99
Gedo_rain	0.35	0%	0.00	100%	0.35
Jubbada_Hoose_goatprice	0.23	0%	0.00	100%	0.23
Nugaal_Conflict	0.17	0%	0.00	100%	0.17
Gedo_FutureRegion	0.15	0%	0.00	100%	0.15
Gedo_BeforeRegion	0.14	100%	0.14	0%	0.00
Jubbada_Dhexe_Conflict	0.13	100%	0.13	0%	0.00
Togdheer_WaterDrumPrice	0.07	100%	0.07	0%	0.00
BA_6 Variables	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.97	100%	0.97	0%	0.00
Mudug_BeforeRegion	0.52	100%	0.52	0%	0.00
Togdheer_BeforeRegion	0.40	0%	0.00	100%	0.40
Bakool_WaterDrumPrice	0.39	100%	0.39	0%	0.00
Galgaduud_BeforeRegion	0.30	100%	0.30	0%	0.00
Sanaag_Fatalities	0.26	100%	0.26	0%	0.00
Gedo_CurrentRegion	0.20	100%	0.20	0%	0.00
Gedo_goatprice	0.18	0%	0.00	100%	0.18
Hiiraan_Bulo_Burti_StationShabelle_River	0.07	100%	0.07	0%	0.00
Galgaduud_FutureRegion	0.03	100%	0.03	0%	0.00
BA_5 Variables	Sensitivity	% Positive	PM	% Negative	NM
Bakool_WaterDrumPrice	0.65	85%	0.59	15%	0.98
Jubbada_Hoose_goatprice	0.20	0%	0.00	100%	0.20
Gedo_BeforeRegion	0.20	100%	0.20	0%	0.00
Mudug_BeforeRegion	0.19	100%	0.19	0%	0.00
Sanaag_Fatalities	0.15	100%	0.15	0%	0.00
Togdheer_Fatalities	0.14	100%	0.14	0%	0.00
Gedo_CurrentRegion	0.11	100%	0.11	0%	0.00
Jubbada_Hoose_FutureRegion	0.10	100%	0.10	0%	0.00
Jubbada_Dhexe_CurrentRegion	0.07	100%	0.07	0%	0.00
Sanaag_CurrentRegion	0.03	0%	0.00	100%	0.03

Table A.10: Sensitivity of Models for September in descending order 7-5

BA_4 Variables	Sensitivity	% Positive	PM	% Negative	NM
Togdheer_CurrentRegion	1.29	100%	1.29	0%	0.00
Togdheer_BeforeRegion	0.57	0%	0.00	100%	0.57
Bay_BeforeRegion	0.31	100%	0.31	0%	0.00
Jubbada_Hoose_goatprice	0.24	0%	0.00	100%	0.24
Sanaag_Conflict	0.24	100%	0.24	0%	0.00
Bakool_WaterDrumPrice	0.21	100%	0.21	0%	0.00
Galgaduud_WaterDrumPrice	0.14	100%	0.14	0%	0.00
Gedo_CurrentRegion	0.11	100%	0.11	0%	0.00
Mudug_BeforeRegion	0.10	100%	0.10	0%	0.00
Nugaal_Conflict	0.08	0%	0.00	100%	0.08
Hiiraan_Conflict	0.04	100%	0.04	0%	0.00
Juba_River_discharge	0.04	100%	0.04	0%	0.00
Gedo_DollowStation_Juba_River	0.02	100%	0.02	0%	0.00
BA_3 Variables	**Sensitivity**	**% Positive**	**PM**	**% Negative**	**NM**
Mudug_FutureRegion	2.33	100%	2.33	0%	0.00
Hiiraan_BeforeRegion	1.89	100%	1.89	0%	0.00
Bakool_WaterDrumPrice	0.55	100%	0.55	0%	0.00
Mudug_BeforeRegion	0.51	100%	0.51	0%	0.00
Galgaduud_FutureRegion	0.45	100%	0.45	0%	0.00
Gedo_CurrentRegion	0.25	100%	0.25	0%	0.00
Bari_CurrentRegion	0.18	100%	0.18	0%	0.00
Shabeellaha_Dhexe_BeforeRegion	0.10	100%	0.10	0%	0.00
BA_2 Variables	**Sensitivity**	**% Positive**	**PM**	**% Negative**	**NM**
Banadir_WaterDrumPrice	1.31	57%	1.28	43%	1.36
Bakool_WaterDrumPrice	1.07	76%	1.08	24%	1.04
Jubbada_Hoose_goatprice	0.23	0%	0.00	100%	0.23
Mudug_BeforeRegion	0.18	100%	0.18	0%	0.00
Gedo_BeforeRegion	0.15	100%	0.15	0%	0.00
Sanaag_Fatalities	0.12	100%	0.12	0%	0.00
Gedo_CurrentRegion	0.11	100%	0.11	0%	0.00
Togdheer_Fatalities	0.08	100%	0.08	0%	0.00
Jubbada_Dhexe_CurrentRegion	0.07	100%	0.07	0%	0.00

Table A.11: Sensitivity of Models for September in descending order 4-2

BA_1 Variables	Sensitivity	% Positive	PM	% Negative	NM
Bakool_rain	0.64	100%	0.64	0%	0.00
Hiiraan_BeforeRegion	0.34	100%	0.34	0%	0.00
Gedo_Fatalities	0.32	100%	0.32	0%	0.00
Mudug_Fatalities	0.28	100%	0.28	0%	0.00
Bakool_WaterDrumPrice	0.25	100%	0.25	0%	0.00
Gedo_CurrentRegion	0.25	100%	0.25	0%	0.00
Mudug_BeforeRegion	0.24	100%	0.24	0%	0.00
Nugaal_CurrentRegion	0.16	100%	0.16	0%	0.00
Bari_Fatalities	0.15	100%	0.15	0%	0.00
Bari_CurrentRegion	0.12	100%	0.12	0%	0.00
Woqooyi_Galbeed_FutureRegion	0.10	100%	0.10	0%	0.00

Table A.12: Sensitivity of Models for September in descending order 1

Variable GROUPED	Sensitivity	Frequency
Gedo_rain	20.75	3
Hiiraan_BeforeRegion	11.50	9
Bay_BeforeRegion	7.35	10
Banadir_BeforeRegion	3.38	10
Mudug_Fatalities	2.59	8
Nugaal_CurrentRegion	1.44	4
Mudug_BeforeRegion	1.37	8
Gedo_BeforeRegion	1.22	3
Bari_Conflict	1.04	7
Sool_CurrentRegion	0.95	2
Hiiraan_WaterDrumPrice	0.84	5
Awdal_Conflict	0.72	2
Shabeellaha_Dhexe_BeforeRegion	0.52	5
Nugaal_FutureRegion	0.36	4
Mudug_CurrentRegion	0.33	1
Galgaduud_FutureRegion	0.30	1
Gedo_FutureRegion	0.27	2
Sanaag_Fatalities	0.27	2
Awdal_BeforeRegion	0.25	1
Galgaduud_CurrentRegion	0.24	1
Togdheer_Fatalities	0.23	2
Bay_WaterDrumPrice	0.19	2
Woqooyi_Galbeed_rain	0.15	1
Gedo_goatprice	0.13	1
Nugaal_Fatalities	0.12	1

Table A.18: Summed Sensitivity of Models for June and Frequency

BA_JUN1 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.40	97%	0.41	3%	0.02
Mudug_Fatalities	0.34	100%	0.34	0%	0.00
Banadir_BeforeRegion	0.30	64%	0.20	36%	0.49
Gedo_BeforeRegion	0.28	100%	0.28	0%	0.00
Hiiraan_BeforeRegion	0.24	100%	0.24	0%	0.00
Hiiraan_WaterDrumPrice	0.20	100%	0.20	0%	0.00
Gedo_FutureRegion	0.19	69%	0.11	31%	0.37
Togdheer_Fatalities	0.16	91%	0.17	9%	0.10
Bari_Conflict	0.12	94%	0.10	6%	0.42
Mudug_BeforeRegion	0.01	100%	0.01	0%	0.00
BA_JUN2 Variable	Sensitivity	% Positive	PM	% Negative	NM
Gedo_rain	7.68	100%	7.68	0%	0.00
Hiiraan_BeforeRegion	0.98	100%	0.98	0%	0.00
Bay_BeforeRegion	0.64	100%	0.64	0%	0.00
Mudug_Fatalities	0.35	100%	0.35	0%	0.00
Mudug_CurrentRegion	0.33	0%	0.00	100%	0.33
Banadir_BeforeRegion	0.25	100%	0.25	0%	0.00
Bari_Conflict	0.17	100%	0.17	0%	0.00
Mudug_BeforeRegion	0.11	100%	0.11	0%	0.00
Shabeellaha_Dhexe_BeforeRegion	0.10	100%	0.10	0%	0.00
Bakool_WaterDrumPrice	0.09	100%	0.09	0%	0.00
Hiiraan_WaterDrumPrice	0.09	100%	0.09	0%	0.00
BA_JUN3 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.62	100%	0.62	0%	0.00
Sool_CurrentRegion	0.48	100%	0.48	0%	0.00
Banadir_BeforeRegion	0.45	100%	0.45	0%	0.00
Awdal_Conflict	0.36	100%	0.36	0%	0.00
Mudug_BeforeRegion	0.28	100%	0.28	0%	0.00
Hiiraan_BeforeRegion	0.28	100%	0.28	0%	0.00
Nugaal_CurrentRegion	0.26	100%	0.26	0%	0.00
Sanaag_Fatalities	0.13	100%	0.13	0%	0.00
Shabeellaha_Dhexe_BeforeRegion	0.12	100%	0.12	0%	0.00
Bay_WaterDrumPrice	0.10	100%	0.10	0%	0.00
Nugaal_FutureRegion	0.03	100%	0.03	0%	0.00

Table A.13: Sensitivity of Models for June in ascending order 1-3

BA_JUN4 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.62	100%	0.62	0%	0.00
Sool_CurrentRegion	0.48	100%	0.48	0%	0.00
Banadir_BeforeRegion	0.45	100%	0.45	0%	0.00
Awdal_Conflict	0.36	100%	0.36	0%	0.00
Mudug_BeforeRegion	0.28	100%	0.28	0%	0.00
Hiiraan_BeforeRegion	0.28	100%	0.28	0%	0.00
Nugaal_CurrentRegion	0.26	100%	0.26	0%	0.00
Sanaag_Fatalities	0.13	100%	0.13	0%	0.00
Shabeellaha_Dhexe_BeforeRegion	0.12	100%	0.12	0%	0.00
Bay_WaterDrumPrice	0.10	100%	0.10	0%	0.00
Nugaal_FutureRegion	0.03	100%	0.03	0%	0.00
BA_JUN5 Variable	Sensitivity	% Positive	PM	% Negative	NM
Gedo_rain	6.60	100%	6.60	0%	0.00
Hiiraan_BeforeRegion	0.83	100%	0.83	0%	0.00
Bay_BeforeRegion	0.78	100%	0.78	0%	0.00
Mudug_Fatalities	0.43	100%	0.43	0%	0.00
Banadir_BeforeRegion	0.25	100%	0.25	0%	0.00
Awdal_BeforeRegion	0.25	0%	0.00	100%	0.25
Hiiraan_WaterDrumPrice	0.18	100%	0.18	0%	0.00
Woqooyi_Galbeed_rain	0.15	0%	0.00	100%	0.15
Bari_Conflict	0.13	100%	0.13	0%	0.00
Jubbada_Dhexe_FutureRegion	0.03	0%	0.00	100%	0.03
Nugaal_WaterDrumPrice	0.02	0%	0.00	100%	0.02
Mudug_BeforeRegion	0.00	100%	0.00	0%	0.00
BA_JUN6 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.61	100%	0.61	0%	0.00
Nugaal_CurrentRegion	0.57	100%	0.57	0%	0.00
Hiiraan_BeforeRegion	0.36	100%	0.36	0%	0.00
Mudug_Fatalities	0.32	100%	0.32	0%	0.00
Bari_Conflict	0.28	100%	0.28	0%	0.00
Banadir_BeforeRegion	0.26	100%	0.26	0%	0.00
Hiiraan_WaterDrumPrice	0.19	100%	0.19	0%	0.00
Gedo_FutureRegion	0.08	100%	0.08	0%	0.00
Togdheer_Fatalities	0.06	100%	0.06	0%	0.00

Table A.14: Sensitivity of Models for June in descending order 4-6

BA_JUN7 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	0.63	100%	0.63	0%	0.00
Banadir_BeforeRegion	0.61	100%	0.61	0%	0.00
Mudug_Fatalities	0.31	94%	0.32	6%	0.11
Mudug_BeforeRegion	0.19	100%	0.19	0%	0.00
Gedo_BeforeRegion	0.15	100%	0.15	0%	0.00
Nugaal_Fatalities	0.12	63%	0.04	38%	0.26
Hiiraan_CurrentRegion	0.10	100%	0.10	0%	0.00
Shabeellaha_Dhexe_BeforeRegion	0.10	100%	0.10	0%	0.00
Bari_Conflict	0.10	100%	0.10	0%	0.00
BA_JUN8 Variable	Sensitivity	% Positive	PM	% Negative	NM
Gedo_rain	6.47	100%	6.47	0%	0.00
Hiiraan_BeforeRegion	4.56	100%	4.56	0%	0.00
Bay_BeforeRegion	0.62	100%	0.62	0%	0.00
Nugaal_CurrentRegion	0.36	100%	0.36	0%	0.00
Mudug_Fatalities	0.29	100%	0.29	0%	0.00
Banadir_BeforeRegion	0.23	100%	0.23	0%	0.00
Hiiraan_WaterDrumPrice	0.19	100%	0.19	0%	0.00
Bari_Conflict	0.13	100%	0.13	0%	0.00
Gedo_CurrentRegion	0.02	100%	0.02	0%	0.00
BA_JUN9 Variable	Sensitivity	% Positive	PM	% Negative	NM
Hiiraan_BeforeRegion	3.90	100%	3.90	0%	0.00
Gedo_BeforeRegion	0.79	100%	0.79	0%	0.00
Bay_BeforeRegion	0.51	100%	0.51	0%	0.00
Mudug_BeforeRegion	0.33	100%	0.33	0%	0.00
Galgaduud_FutureRegion	0.30	100%	0.30	0%	0.00
Mudug_Fatalities	0.27	100%	0.27	0%	0.00
Banadir_BeforeRegion	0.21	100%	0.21	0%	0.00
Gedo_goatprice	0.13	0%	0.00	100%	0.13
Bari_Conflict	0.11	100%	0.11	0%	0.00
Nugaal_FutureRegion	0.02	100%	0.02	0%	0.00

Table A.15: Sensitivity of Models for June in ascending order 7-9

BA_JUN10 Variable	Sensitivity	% Positive	PM	% Negative	NM
Bay_BeforeRegion	1.93	100%	1.93	0%	0.00
Banadir_BeforeRegion	0.37	100%	0.37	0%	0.00
Nugaal_FutureRegion	0.29	100%	0.29	0%	0.00
Mudug_Fatalities	0.27	100%	0.27	0%	0.00
Galgaduud_CurrentRegion	0.24	0%	0.00	100%	0.24
Mudug_BeforeRegion	0.16	100%	0.16	0%	0.00
Hiiraan_Belet_WeyneStation_Shabelle_River	0.12	100%	0.12	0%	0.00
Shabeellaha_Dhexe_rain	0.09	0%	0.00	100%	0.09
Shabeellaha_Dhexe_BeforeRegion	0.08	100%	0.08	0%	0.00
Hiiraan_BeforeRegion	0.07	100%	0.07	0%	0.00
Bari_CurrentRegion	0.07	100%	0.07	0%	0.00
Bakool_FutureRegion	0.05	100%	0.05	0%	0.00
Gedo_DollowStation_Juba_River	0.01	100%	0.01	0%	0.00

Table A.16: Sensitivity of Models for June in ascending order 10

Variable GROUPED	Sensitivity	Frequency
Bakool_WaterDrumPrice	9.54	9
Mudug_BeforeRegion	5.00	10
Mudug_FutureRegion	2.63	2
Hiiraan_BeforeRegion	2.55	4
Gedo_CurrentRegion	1.85	9
Bari_FutureRegion	1.46	2
Galgaduud_FutureRegion	1.36	4
Banadir_WaterDrumPrice	1.31	1
Togdheer_CurrentRegion	1.29	1
Bay_BeforeRegion	1.29	2
Bari_CurrentRegion	1.11	4
Jubbada_Hoose_goatprice	1.06	5
Togdheer_BeforeRegion	0.97	2
Sanaag_Fatalities	0.85	4
Shabeellaha_Dhexe_Fatalities	0.78	2
Gedo_BeforeRegion	0.67	4
Bakool_rain	0.64	1
Jubbada_Dhexe_CurrentRegion	0.61	3
Gedo_goatprice	0.40	2
Nugaal_CurrentRegion	0.38	2
Gedo_rain	0.35	1
Gedo_Fatalities	0.32	1
Galgaduud_BeforeRegion	0.30	1
Bari_rain	0.28	1
Mudug_Fatalities	0.28	1
Nugaal_Conflict	0.25	2
Sanaag_Conflict	0.24	1
Togdheer_Fatalities	0.23	2
Bari_Fatalities	0.15	1
Shabeellaha_Hoose_rain	0.15	1
Gedo_FutureRegion	0.15	1
Galgaduud_WaterDrumPrice	0.14	1
Jubbada_Dhexe_Conflict	0.13	1
Bakool_CurrentRegion	0.13	1
Jubbada_Hoose_BeforeRegion	0.12	1
Bay_WaterDrumPrice	0.12	1
Jubbada_Hoose_FutureRegion	0.10	1
Shabeellaha_Dhexe_BeforeRegion	0.10	1
Woqooyi_Galbeed_FutureRegion	0.10	1
Nugaal_FutureRegion	0.08	2

Table A.17: Summed Sensitivity of Models for September and Frequency

A.3 REGRESSION TECHNIQUES

ROUND 1 Linear	actual	predicted	err
2017-07-01	39219	84828	45609
2017-08-01	25768	4395	18164
2017-09-01	21554	-23115	-44671
2017-10-01	18461	53888	35427
2017-11-01	24302	50074	25772
2017-12-01	44009	52909	8900
2018-01-01	44926	6312	-38614
2018-02-01	36822	108481	71659
2018-03-01	115474	217382	101908
2018-04-01	47045	132516	85471

ROUND 1 NN	actual	predicted	err
2017-07-01	39219	118295	79076
2017-08-01	25768	92972	67204
2017-09-01	21554	32891	11337
2017-10-01	18461	184111	165650
2017-11-01	24302	113472	89170
2017-12-01	44009	119991	75982
2018-01-01	44926	102376	57450
2018-02-01	36822	115451	78629
2018-03-01	115474	67932	-47542
2018-04-01	47045	137091	90046

Table A.19: ROUND 1 Predictions and Comparison to Actual Arrivals

ROUND 1 Linear	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	47619	157	213	126	55426	3072052497
2-step-ahead	9	47843	157	213	127	56412	3182261893
3-step-ahead	8	51553	207	255	134	59488	3538804916
4-step-ahead	7	52536	195	234	124	61313	3759281730
5-step-ahead	6	72800	197	258	158	79423	6308048661
6-step-ahead	5	76980	183	242	147	83851	7030974968
7-step-ahead	4	99858	217	287	175	107455	11546590369
8-step-ahead	3	99127	207	253	158	111396	12409113564
9-step-ahead	2	115389	238	237	131	133028	17696561001
10-step-ahead	1	49194	629	629	105	49194	2420007241

Table A.20: Round 1 Linear

ROUND 2 Linear	actual	pred	error
2017-07-01	39219	7968.2973	-31250.7027
2017-08-01	25768	18648.9503	-7119.0497
2017-09-01	21554	23219.6201	1665.6201
2017-10-01	18461	31834.0332	13373.0332
2017-11-01	24302	4356.253	19177.253
2017-12-01	44009	7968.9709	-36040.0291
2018-01-01	44926	45533.4611	607.4611
2018-02-01	36822	16491.2656	-20330.7344
2018-03-01	115474	1071.0943	-114402.905
2018-04-01	47045	32017.5162	-15027.4838

ROUND 2 NN	actual	predicted	error
2017-07-01	39219	26942.2376	-12276.7624
2017-08-01	25768	39740.0176	13972.0176
2017-09-01	21554	28715.4903	7161.4903
2017-10-01	18461	59268.7311	40807.7311
2017-11-01	24302	29756.928	5454.928
2017-12-01	44009	6233.9953	-37775.0047
2018-01-01	44926	29894.7413	-15031.2587
2018-02-01	36822	21754.394	-15067.606
2018-03-01	115474	15837.8478	-99636.1522
2018-04-01	47045	40253.3848	-6791.6152

Table A.21: ROUND 2 Predictions and Comparison to Actual Arrivals

ROUND 1 NN	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	76209	238	337	253	84802	7191409117
2-step-ahead	9	108070	332	481	389	118949	14148983073
3-step-ahead	8	134305	518	664	481	148588	22078381721
4-step-ahead	7	187488	659	835	619	206955	42830225306
5-step-ahead	6	226638	586	804	592	236204	55792240800
6-step-ahead	5	236693	519	744	500	238381	56825413487
7-step-ahead	4	207255	420	595	420	207636	43112698805
8-step-ahead	3	191800	361	489	389	194443	37807950942
9-step-ahead	2	172713	315	354	281	176307	31084126404
10-step-ahead	1	163641	2091	2091	348	163641	26778433748

Table A.22: Round 1 NN

ROUND 2 Linear	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	25899	116	113	54	40747	1660310819
2-step-ahead	9	25332	116	113	51	41690	1738097593
3-step-ahead	8	27665	154	137	54	44169	1950859251
4-step-ahead	7	31452	150	140	60	47217	2229423364
5-step-ahead	6	34483	126	122	59	50710	2571486493
6-step-ahead	5	37538	119	118	55	54882	3011983274
7-step-ahead	4	37974	119	109	48	58691	3444635123
8-step-ahead	3	50340	126	128	63	67769	4592591788
9-step-ahead	2	65043	146	133	66	81669	6669754047
10-step-ahead	1	15655	200	200	33	15655	245066469

Table A.23: ROUND 2 Linear

ROUND 2 NN	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	25397	109	119	62	37344	1394566161
2-step-ahead	9	28375	111	126	70	39670	1573719459
3-step-ahead	8	30786	147	152	74	42240	1784213723
4-step-ahead	7	34215	144	152	80	45097	33722673
5-step-ahead	6	32873	113	117	55	45549	2074696067
6-step-ahead	5	37521	108	118	58	49706	2470645248
7-step-ahead	4	37711	106	108	51	52444	2750407762
8-step-ahead	3	45049	111	115	57	59875	3585070881
9-step-ahead	2	58819	129	121	62	72279	5224323739
10-step-ahead	1	16811	215	215	36	16811	282600203

Table A.24: ROUND 2 NN

ROUND 3 Linear	actual	predicted	error
2017-07-01	39219	31587.3535	-7631.6465
2017-08-01	25768	41833.2536	16065.2536
2017-09-01	21554	14631.5535	-6922.4465
2017-10-01	18461	38448.8034	19987.8034
2017-11-01	24302	46798.4641	22496.4641
2017-12-01	44009	9041.7299	-34967.2701
2018-01-01	44926	15922.312	-29003.688
2018-02-01	36822	36929.2069	107.2069
2018-03-01	115474	54290.2229	-61183.7771
2018-04-01	47045	60546.8759	13501.8759

ROUND 3 NN	actual	predicted	error
2017-07-01	39219	30900.6268	-8318.3732
2017-08-01	25768	27986.1455	2218.1455
2017-09-01	21554	30948.7038	9394.7038
2017-10-01	18461	1466.1383	-16994.8617
2017-11-01	24302	1422.2377	-22879.7623
2017-12-01	44009	1421.6213	-42587.3787
2018-01-01	44926	1435.6043	-43490.3957
2018-02-01	36822	5689.4304	-31132.5696
2018-03-01	115474	17342.8237	-98131.1763
2018-04-01	47045	1435.0189	-45609.9811

Table A.25: ROUND 3 Predictions and Comparison to Actual Arrivals

ROUND 3 Linear	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	21187	79	101	54	26944	725954846
2-step-ahead	9	22636	79	101	58	28268	799064727
3-step-ahead	8	23467	103	116	57	29445	866980103
4-step-ahead	7	25881	100	115	61	31554	995663051
5-step-ahead	6	26625	81	94	53	32756	1072951560
6-step-ahead	5	27414	75	86	44	34479	1188828675
7-step-ahead	4	26007	70	75	37	34680	1202675252
8-step-ahead	3	24838	67	63	27	36235	1313001844
9-step-ahead	2	36538	79	75	39	44370	1968699721
10-step-ahead	1	11361	145	145	24	11361	129068601

Table A.26: ROUND 3 Linear

ROUND 3 NN	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	32076	122	154	72	41664	1735901919
2-step-ahead	9	34555	122	154	77	43563	1897722945
3-step-ahead	8	38387	160	190	86	45887	2105604695
4-step-ahead	7	42394	155	189	91	48709	372577942
5-step-ahead	6	46573	129	165	91	52077	2712062616
6-step-ahead	5	51079	121	161	91	55567	3087744005
7-step-ahead	4	52996	117	152	89	58064	3371467503
8-step-ahead	3	56165	115	143	86	62169	3864930378
9-step-ahead	2	69190	131	142	89	73102	5343968607
10-step-ahead	1	45593	583	583	97	45593	2078750329

Table A.27: ROUND 3 NN

A.3.1 ROUND 4

LR COMMON 2	actual	pred	err
Jul , 2017	39219	35309	-3910
Aug , 2017	25768	74776	49008
Sep , 2017	21554	28366	6812
Oct , 2017	18461	34378	15917
Nov , 2017	24302	-10045	-34347
Dec , 2017	44009	4192	-39817
Jan , 2018	44926	25945	-18981
Feb , 2018	36822	70271	33449
Mar , 2018	115474	53317	-62157
Apr , 2018	47045	32059	-14986

Table A.28: LR COMMON 2

LR COMMON 2	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	27938	98	136	77	33240	1104907034
2-step-ahead	9	29089	98	129	79	35220	1240475586
3-step-ahead	8	26531	115	131	65	33025	90635084
4-step-ahead	7	30207	112	134	74	35315	247151396
5-step-ahead	6	32453	93	115	72	37534	1408830420
6-step-ahead	5	31697	82	100	56	37790	28092014
7-step-ahead	4	30795	78	88	50	38386	73482503
8-step-ahead	3	35793	81	91	55	43376	81466038
9-step-ahead	2	36492	84	75	36	47231	30722178
10-step-ahead	1	6509	83	83	14	6509	42360799

Table A.29: LR COMMON 2 METRICS

NN COMMON 2	actual	pred	err
Jul , 2017	39219	22994	-16225
Aug , 2017	25768	83248	57480
Sep , 2017	21554	46429	24875
Oct , 2017	18461	27638	9177
Nov , 2017	24302	4517	-19785
Dec , 2017	44009	2418	-41591
Jan , 2018	44926	11731	-33195
Feb , 2018	36822	18220	-18602
Mar , 2018	115474	20580	-94894
Apr , 2018	47045	44344	-2701

Table A.30: NN COMMON 2

NN COMMON 2	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	31852	120	149	82	41051	1685149728
2-step-ahead	9	36178	123	161	93	44019	1937652677
3-step-ahead	8	35246	152	174	82	43634	1903961779
4-step-ahead	7	36231	145	161	75	45487	2069095033
5-step-ahead	6	39885	121	142	74	48803	2381769435
6-step-ahead	5	44137	115	139	73	52807	2788625677
7-step-ahead	4	44767	112	129	68	55252	3052758816
8-step-ahead	3	47640	112	122	64	60288	3634694442
9-step-ahead	2	59330	128	122	63	71818	5157835772
10-step-ahead	1	18860	241	241	40	18860	355712884

Table A.31: NN COMMON 2 METRICS

LR COMMON 6	actual	pred	err
Jul , 2017	39219	49310	10091
Aug , 2017	25768	16260	-9508
Sep , 2017	21554	-9900	-31454
Oct , 2017	18461	5254	-13207
Nov , 2017	24302	24929	627
Dec , 2017	44009	2499	-41510
Jan , 2018	44926	6219	-38707
Feb , 2018	36822	25348	-11474
Mar , 2018	115474	7199	-108275
Apr , 2018	47045	54362	7317

Table A.32: LR COMMON 6

LR COMMON 6	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step-ahead	10	27217	119	129	60	40603	1648640150
2-step-ahead	9	29116	119	129	64	42763	1828652404
3-step-ahead	8	31520	157	156	68	45100	2034030330
4-step-ahead	7	31619	150	141	56	47268	2234237010
5-step-ahead	6	34913	127	124	55	51097	2610911629
6-step-ahead	5	41882	122	132	65	56124	3149897924
7-step-ahead	4	41847	120	120	57	59500	3540227911
8-step-ahead	3	41960	120	107	45	64437	4152122755
9-step-ahead	2	56219	140	115	50	78320	6133998220
10-step-ahead	1	1741	22	22	4	1741	3032711

Table A.33: LR COMMON 6 METRICS

NN COMMON 6	actual	pred	err
Jul , 2017	39219	76895	37676
Aug , 2017	25768	578724	552956
Sep , 2017	21554	4706779	4685225
Oct , 2017	18461	87743	69282
Nov , 2017	24302	59758	35456
Dec , 2017	44009	25231	-18778
Jan , 2018	44926	20954	-23972
Feb , 2018	36822	46568	9746
Mar , 2018	115474	29749	-85725
Apr , 2018	47045	74354	27309

Table A.34: NN COMMON 6

NN COM 6	N	MAE	RRSE	RAE	MAPE	RMSE	MSE
1-step	10	554612	4388	2721	2476	1492437	2227368828486
2-step	9	3382389	19719	15040	17061	7069018	49971019388701
3-step	8	6043608	34499	29873	28535	9897211	97954793474495
4-step	7	7419714	35995	33029	29168	11306025	127826203327833
5-step	6	8656384	30289	30715	24802	12216301	149238019970752
6-step	5	10381294	29128	32651	24937	13380835	179046751449572
7-step	4	12938103	30174	37161	25061	14929801	222898966158869
8-step	3	17334135	32184	44222	32979	17334173	300473544653040
9-step	2	17350346	30994	35579	25974	17350380	301035683646299
10-step	1	17384561	222139	222139	36953	17384561	302222950485934

Table A.35: NN COMMON 6 METRICS