UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

A TIME SERIES FORECASTING APPROACH TO PREDICTION OF REFUGEE POPULATION IN KENYA

BY

MAINGI, SAMUEL KIMANI

(P54/79168/2015)

Supervisor

DR. A. WAUSI

A project report submitted in partial fulfilment of the requirement for the award of Masters of Science Degree in Information Technology Management of the University of Nairobi.

2016

DECLARATION

This project is my original work and to the best of my knowledge this research work has not been submitted for any other award in any university.

Signed :……………………………… Date :………………………………

Maingi Samuel Kimani.

P54/79168/2015

This project report has been submitted in partial fulfillment of requirement for the Master of Science Degree in Information Technology Management of the University of Nairobi with my approval as the University supervisor.

Signed :……………………………… Date :………………………………

Dr. A.N. Wausi

ABSTRACT

Kenya has for many years been referred to as the Oasis of Peace in the Sea of Instability, this is due to the country's geographical and political position as a safe haven for refugees fleeing from war and civil unrest from neighboring countries of Somalia, Uganda, Rwanda, Burundi, South Sudan, Sudan and Democratic Republic of Congo. By the end of December 2015 the total number of refugees and asylum seekers in Kenya was close to 600,000, more than half of that population arrived between 2011 and 2014 alone. Management of refugee affairs requires a diligent approach to availing resources when required (during a refugee emergency) or relinquishing resources when numbers decrease, both actions of response cannot be achieved without an effective methodology for predicting future refugee population.

This paper uses trend analysis using Weka data mining software. The data used for this project was extracted from the population data reported by the United Nations High Commissioner for Refugees (UNHCR) in the organization's public statistics portal. The methodology used for implementing this project is CRISP-DM which guided the extraction and preparation of data to feeding it to the data mining tool. A number of algorithms were tested for against the historical data and Multilayer Perceptron which is based on regression was found to be most preferable for creation of prediction models.

The study revealed that the models created through training on the historical data performed quite well. The end of year 2016 prediction was compared against the population figures reported at the end of September 2016 and they were found to be within acceptable range. One aspect that was found to affect the prediction was uncertainty, the study revealed that drastic changes in refugee population such as through outbreak of violence in the refugee countries of origin could not be captured by the trained models therefore the use of expert input adopted to complement the trend learned by the models.

The paper concluded by providing recommendations which emphasized on the use of a hybrid approach in handling uncertainty through the use of the expert input and probability functions to calculate the likelihood of selected factors impacting the population as proposed by the experts. In addition the research recommended a sectoral approach to population prediction to look at aspects like prediction of malnutrition rates, health, education etc. this would require consistent collection and storage of such information so that it can be used to train prediction models.

ACKNOWLEDGEMENTS

This project has been produced with contributions from Dr. A.N Wausi (Supervisor), Dr. E. Opiyo, Dr. R.O. Oboko, Ms. P. W. Wangunyu all from the School of Computing and Informatics, University of Nairobi.

I would not forget to give thanks to The Almighty God for the unremitting favors He has bestowed upon me throughout the project.

Contents

List of Tables

List of Figures

List of Abbreviations and Definitions

i.	UNHCR	The United Nations High Commissioner for Refugees
ii.	DRA	Government of Kenya Department of Refugee Affairs
iii.	DM	Data Mining
iv.	Refugee	A refugee is someone who has been forced to flee his or her country because of persecution, war, or violence. A refugee has a well-founded fear of persecution for reasons of race, religion, nationality, political opinion or membership in a particular social group. Most likely, they cannot return home or are afraid to do so. War and ethnic, tribal and religious violence are leading causes of refugees fleeing their countries.
v.	Asylum seeker	When people flee their own country and seek sanctuary in another country, they apply for asylum – the right to be recognized as a refugee and receive legal protection and material assistance. An asylum seeker must demonstrate that his or her fear of persecution in his or her home country is well-founded.
vi.	RMSE	Root Mean Squared Error
vii.	MAPE	Mean Absolute Percentage Error
viii.	CRISP-DM	Cross-Industry Standard Process for data mining

CHAPTER 1: INTRODUCTION

1.1. Background

For more than thirty years, Kenya has played host to a number of refugees and asylum seekers fleeing from their countries due to a well-founded fear of persecution. (UNHCR country operations profile – Kenya, 2015). The main countries of origin for refugees and asylum seekers in Kenya include the following among others: Somalia, South Sudan, Ethiopia, Congo DRC, Sudan and most recently increasing asylum claims have been received from Burundi. By the end of December 2015 the total number of refugees and asylum seekers in Kenya was close to 600,000, more than half of that population arrived between 2011 and 2014 alone (UNHCR data repository, http://popstats.unhcr.org/en/persons_of_concern).

The formal proposition of the refugee law in Kenya in 1984 was followed by a refugee influx from Uganda, the Government of Kenya therefore became responsible for receiving refugees arriving at its borders. Due to difficult economic times in Kenya in 1992 (inflation rose by more than 50%), the Government handed over the management of refugee affairs to the UNHCR. Law makers continued work on a Refugee Bill and in 2007 the Refugee Act was gazetted which created the Department for Refugee Affairs (DRA) with the mandate to manage all refugee related issues in Kenya. Until its disbandment on 11th May 2016 DRA worked closely with UNHCR which has the international mandate to protect refugees and asylum seekers, the DRA was replaced by a Refugee Affairs Secretariat with a similar mandate.

Refugee operations in Kenya are divided into three, namely Dadaab operation, Kakuma camps and the urban operation in Nairobi. Management of a refugee situation requires a lot of domain knowledge, preparedness and logistics both by government and humanitarian actors especially during a refugee influx. Among the aspects which need to be considered include emergency shelter, food and sanitation as well as registration and enumeration of individuals.

1.2. Problem statement

Readiness for a humanitarian emergency is the single most important aspect of emergency management. Refugees flee from their countries for various reasons such as conflict or war, famine and other conflict induced reasons such as the lack of basic needs and livelihoods. To respond effectively to a refugee emergency, it is important to be able to predict the start of an influx. Knowledge of the start of a refugee influx begins from an understanding of the triggers of flight or reasons which cause individuals to flee from their country of habitual residence.

Among profile related information that would be useful includes the country of origin, the demographics of the individuals seeking asylum as well as the specific area of origin.

The purpose of this project is to establish the existence of patterns of refugee population in Kenya by performing trend analysis on past refugee population data; this is aimed at establishing whether it would be possible to use such patterns to forecast population or predict a recurrence of refugee emergencies.

1.3. Research objectives

The aim of the proposed research is to use data mining techniques to train a model which can use historical data to predict future refugee population in Kenya. This will be achieved through conducting analysis on historical refugee registration data. This project will help the stakeholders managing refugee affairs here in the country to effectively budget and plan for resources over a period.

The following are the main objectives of this study:

i. Training time series forecasting models from historical refugee data.
ii. Using the trained models to make population forecasts

1.4. Research questions

This research will seek to answer the following main question:

> *"Is it possible to use data mining techniques to forecast refugee population trends?"*

1.5. Significance of the study

Statistics of registered refugees and asylum seekers has for a long time been used purely for reporting purposes to take stock at the end of a reporting year. Every year, organizations working in refugee operations use the end year figures to plan budgeting and operations for subsequent years. This research project will help humanitarian organizations as well as the Government of Kenya to understand the history of refugee arrivals and use it to forecast and plan for future needs in response to refugee situations.

1.6. Assumptions of the research

Refugees are said to flee from their country of origin due to various reasons, the reasons that caused refugees to flee in the past would not necessarily be the reasons that cause the flight today or in the future however the overarching reason is the failure of their governments to protect them and provide them with basic needs. It is assumed that this prototype will attempt to learn from the past and attempt to predict the situation at a time in the future. Some situations however cannot be predicted through mathematical methods such as government policy to repatriate refugees which can lead to reduction of a refugee population or a sudden crisis in a country leading to a refugee influx. This limitation of mathematical forecasting was evident during the Global Financial Crisis in 2008 (Bezemer, 2009).

1.7. Scope of the study

The study is based on refugee population data reported by UNHCR Kenya in their public statistics portal (http://popstats.unhcr.org/). The data covers three locations Dadaab, Kakuma and Nairobi and is reported every end of year. This data will be processed and fed into a data mining tool to which will attempt to analyze the trend and provide the output. This study will not go to the details of demographic data elements of different population groups however population forecasting will be attempted per refugee nationality since different countries have different causes for refugee situations.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Hand (2001) defines Data Mining as "The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner". The amount of data in the world today is ever increasing and indeed overwhelming (Witten et. al., 2011). As outlined in Oracle.com, the key properties of data mining are: (1) automatic discovery of patterns, (2) prediction of likely outcomes, (3) creation of actionable information and (4) focus on large data sets and databases. These four properties advised the decision to use data mining to handle the defined problem.

Focus on data mining is mainly in finance and business which see an enormous amount of stock exchange, banking, online and onsite purchasing data flowing daily through computer systems (Halevi & Moed, 2012). The two researchers add that in these fields, the data is captured and stored for inventory monitoring, customer behavior and market behavior; they focused their research on the disciplines which have been making use of business intelligence as well as their geographical distribution and came up with the conclusion summarized by the figures 1 and 2 below.

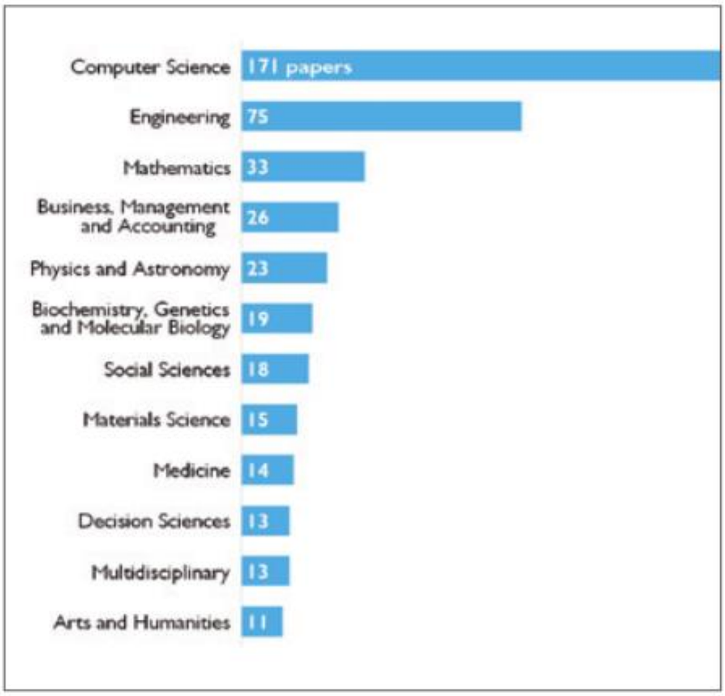


Figure 1: Subject areas researching data mining (Source: Halevi & Moed, 2012)

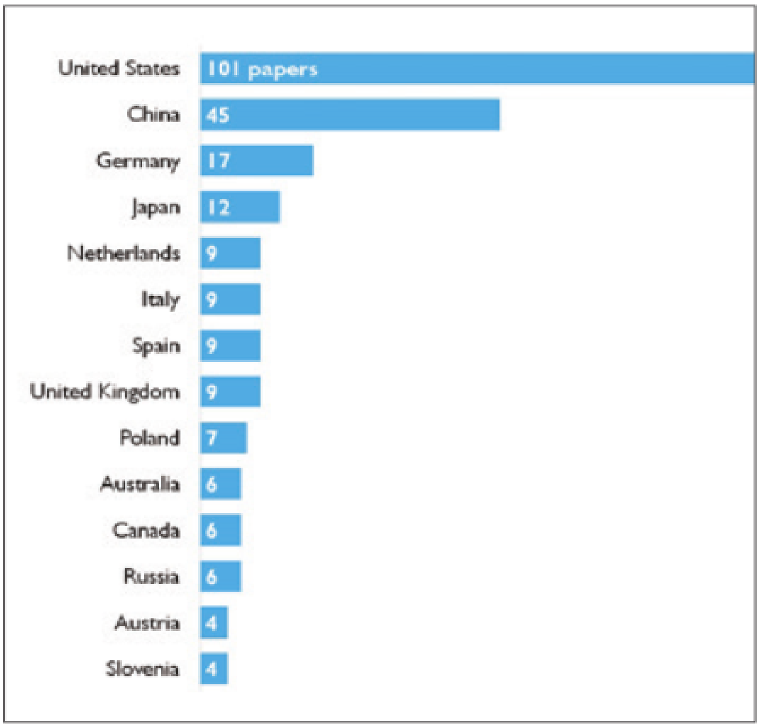


Figure 2: Geographical Distribution of DM papers. (Source: Halevi & Moed, 2012)

From the findings above, the researchers show that there has been little interest in the data mining research in humanities, additionally the geographical distribution of research in data by the year 2012 was not featuring any African countries. In the online article "Big Data in Africa: Its Meaning" (Alliy, 2014), the writer outlines the steps taken by Africa towards adoption of Big Data, however also questions whether Africa is starting to generate these amounts of data in the first place. On the contrary, an article on www.techabal.com states "Africa may trail the US and Europe in terms of technology, but the gap is closing fast", the writer continues to state the reason for increasing use of technology is the zeal and innovation inspired by the economic woes.

In a journal article "Advancing big data for humanitarian needs" (Fadiya et.al, 2014), acknowledge the efforts made by the public, government and humanitarian organizations to adapt to the increasing amount of data available for scientific analysis.

2.2 Context of humanitarian operations

"The social contribution of conflicts, political violence and disasters are all digitized through the creation of data streams" (Fadiya, 2014). Finding ways of increasing humanitarian services with data is very important. In the 2015 article "Guidance for Incorporating Big Data into Humanitarian Operations" the writers Whipkey and Verity concur with Fadiya by stating that big data in humanitarian response is becoming more prominent and important. The writers continue by stating the time and resource constraints of humanitarian organizations and responders continues to become stretched with the growing number and length of humanitarian crises. In addition the writers Whipkey and Verity suggest that what makes many organizations resist incorporating big data practices into response is the constraints mentioned above in combination with the vulnerability of the affected populations which often are the subjects of the data.

2.3 Data factors in humanitarian operations

The United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA) is mandated to co-ordinate humanitarian actors to ensure coherent response during emergencies. Among their functions include the management of information during an emergency. In a June 2013 article in the UNOCHA website called "Big data and humanitarianism: 5 things you need to know", the writers mention some important points to note regarding data and humanitarianism, two of the points highlighted below were found to be relevant to this study:

i. "Finding ways to make data useful to humanitarian decision makers is one of the great challenges, and opportunities, of the network age"
ii. "Data should complement existing sources of information, not replace them"

Akkihal (2006) did some work on inventory prepositioning for humanitarian operations and outlined some important requirements for determining the optimal positions and configurations of inventory for humanitarian emergency response; at the center of response was information about the kind of response required as well as data on the number of individuals affected and their demographic breakdown.

Kenya Inter-Agency Rapid Assessment Mechanism (KIRA)

The Kenya Inter-Agency Rapid Assessment Mechanism (KIRA) is an innovative coordinated needs assessment tool developed by the humanitarian community in Kenya. KIRA was developed around the humanitarian issues surrounding Kenya including the Drought Crisis at the Horn of Africa in 2011 and the post-election violence in Kenya in December 2007 following the disputed presidential elections. At the heart of KIRA's operation is the use of data to guide humanitarian response. The figure below shows the main elements of rapid assessment as designed by KIRA.

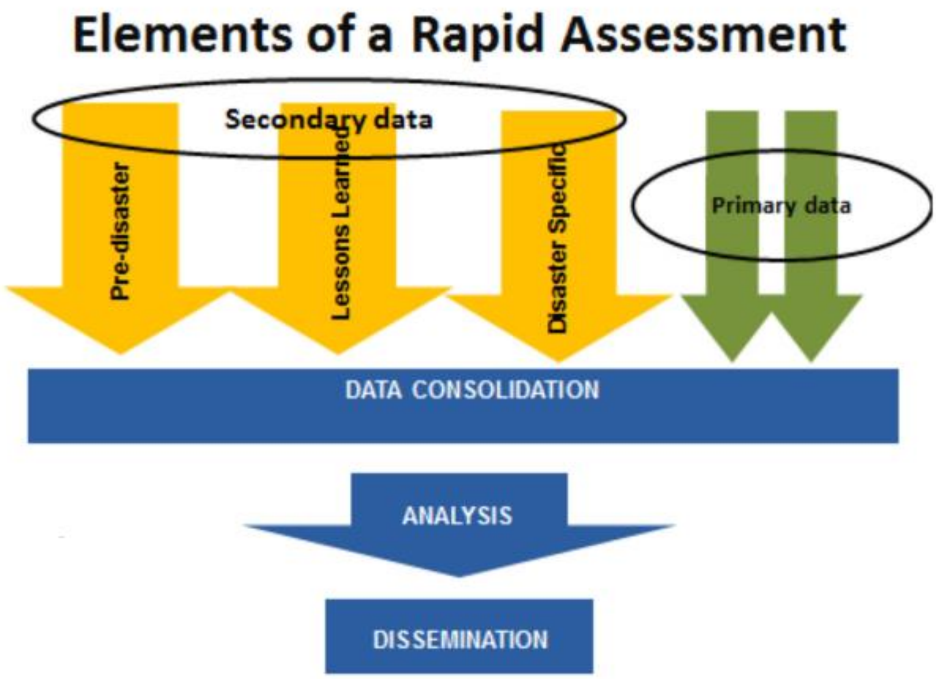


Figure 3: Different elements of KIRA methodology (Limbu et.al, 2015)

The paper by Limbu et.al (2015) emphasizes the importance of data in humanitarian response, it was noted that immediately after the introduction of the data assessment mechanisms, response moved at a faster pace.

2.4 Process Methodologies for Data Mining

A number of methodologies for data mining have been developed and used over the years. These are guides which introduce a structure in the implementation of data mining projects to ensure a goal oriented approach. The following are some of the methodologies employed in various sectors, they were assessed for suitability in the context of this project.

2.4.1 Six Sigma Methodology

This is a data driven methodology for ensuring little or no quality control problems, defects or waste in manufacturing, service and other business related activities. It is characterized by the following steps:

Define → Measure → Analyze → Improve → Control

Often the above steps follow sequentially in a cyclic format. Tamboli (2010), a proponent of the Six Sigma methodology states that the methodology is famous for its data centered approach. He continues to state “The Measure and Analyze phases help to identify why things are the way they are, this knowledge in turn can be used to establish linkages between inputs and outputs; these identified linkages can then help to carry out improvements."

Six sigma methodology has a wide range of application areas including some outside data mining such as in management, manufacturing and business, the flexibility in application areas is a big advantage of this methodology however its lack of specialization could also be seen as a weakness.

2.4.2 SEMMA Methodology

This stands for Sample, Explore, Modify, Model and Assess which is a list of sequential steps developed by SAS Institute Inc. It is considered a general data mining methodology. Like Six Sigma, SEMMA is implemented in a cyclic manner. SEMMA focuses more on the technical activities involved in a typical DM project. The figure below shows the schematic of SEMMA as depicted by the SAS Institute.

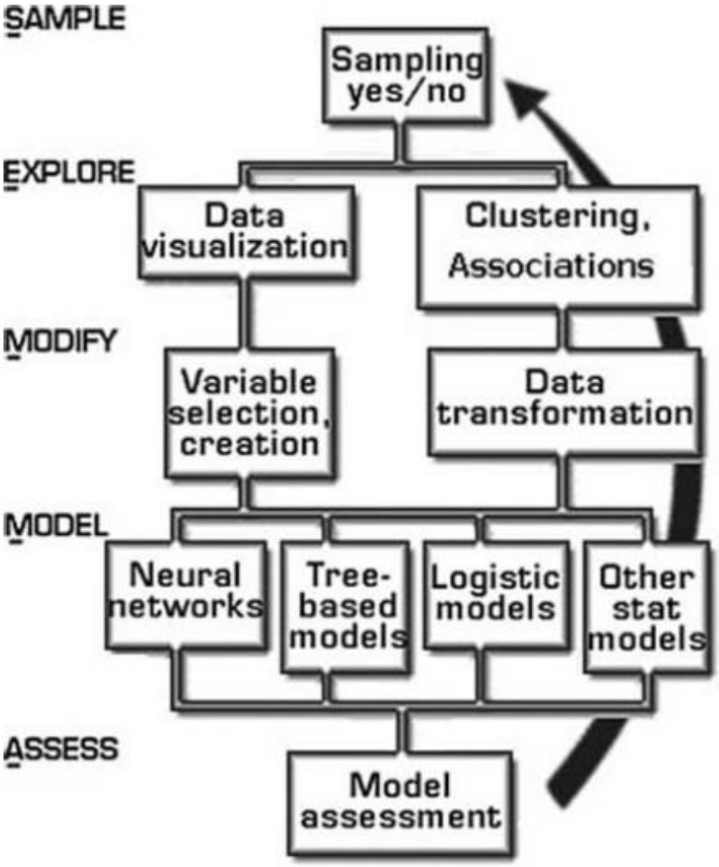


Figure 4: Schematic of SEMMA (SAS Institute)

Rohanizadeh and Moghadam (2009) critique the model by stating that "Although the SEMMA methodology contains some of the essential elements of any data-mining project, it concerns only the statistical, the modeling, and the data manipulation parts of the data-mining process. It lacks some of the fundamental parts of any information systems project, including analysis, design, and implementation phases."

2.4.3 CRISP-DM

This describes common approaches that DM experts use to handle problems. According to IBM, one of the strong points of CRISP-DM is analytics. CRISP-DM contains the following six steps:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

The six phases provided by CRISP-DM are flexible allowing for movement back and forth between different phases. In addition CRISP-DM introduces a user manual guiding data mining researchers' on the required actions and outputs at each stage.

2.4.4 Preferred Process Model (CRISP-DM)

"CRISP-DM was conceived in late 1996 by three *veterans* of the young and immature data mining market. DaimlerChrysler (then Daimler-Benz) was already ahead of most industrial and commercial organizations in applying data mining in its business operations" (CRISP-DM Consortium, 2000). It contains the phases of a project, their respective tasks, and the relationships between these tasks. The reference model defines the following life cycle of a DM project consisting of six phases, shown in Figure below.

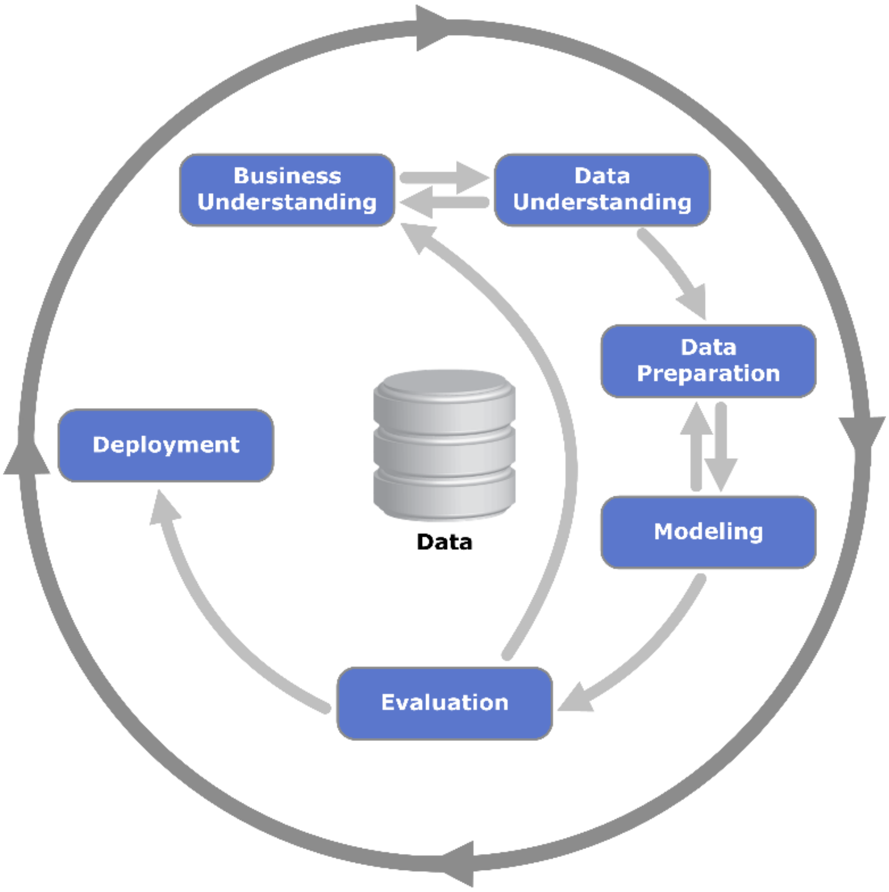


Figure 5: CRISP Data Mining Process Model

The sequence of the phases allows for movement back and forth in a flexible manner as required. The phases of the reference model are as below:

i. *Business understanding*

Here, focus is on understanding the objectives of the project as well as what it requires from a business perspective, the knowledge is then converted into a data mining problem.

ii. *Data understanding*

This starts with initial data collection followed by activities which will allow the researcher to understand the data.

iii. *Data preparation*

This phase will cover the activities necessary to construct the final dataset to be fed into the modeling tools.

iv. *Modeling*

Various techniques of modelling will be selected and applied in this phase, their parameters are calibrated to achieve the most optimal outcome.

v. *Evaluation*

From a data analysis perspective by this stage the researcher will have built a model with high quality, the evaluation phase is done before deployment to find out if there are any pertinent business issues that remain unresolved.

vi. *Deployment*

The proponents of this model suggest that this phase can range from a simple report to a complex repeatable data mining process implemented across an organization.

Rohanizadeh and Moghadam (2009) identified a problem with the approach suggested by CRISP-DM, they state: "The selection of the technique is delayed until the modeling phase; if the data required is not available or is in the wrong format, the model has to return to the data analysis phase again. CRISP-DM, indeed, stresses in the importance of assessing tools and techniques early in the process but also affirms that the selection of tools may influence the entire project."

CRISP-DM was selected for this project, the methodology provides a structured approach to planning a data mining project, it is a robust and well-proven methodology; in addition CRISP-DM is an open methodology allowing for the use of any technology. The six phases provided by CRISP-DM are flexible as illustrated in figure above, allowing for movement back and forth between different phases; this will ensure that the output of the project can be fine-tuned as much as possible to match the objectives. The description of the model to a researcher is very clear in terms of what actually has to be done.

2.5 Predictive Data Mining

The main data mining task in this project will be predictive modelling, the aim of this will be to build a model that will allow the value of a variable to be predicted. Edelstein (1999) distinguishes between two broad types of prediction; Classification predicts into what category a class or case falls while Regression predicts what number value a variable will have and if it varies with time as is the case of this proposed project then it is referred to as a 'time series' prediction. This project will build its model around time series forecasting to predict refugee population in Kenya.

Time series forecasting

Time series forecasting (TSF) predicts values in an unknown future based on a series of predictors varying over time. Similar to regression, TSF uses known results in generating its predictions. A time series can be continuous or discrete (Agrawal, 2013). In a continuous time series observations are measured at every instance of time, whereas a discrete time series contains observations measured at discrete points of time.

Algorithms

An algorithm in is a set of calculations that creates a model from data. The algorithm first analyzes the data provided to create a model while searching for specific types of trends of patterns. The results of this analysis are used by the algorithm uses the results of this analysis over many cycles to find the optimal parameters (to be applied across the entire dataset) for creating the mining model.

2.6 Time Series Forecasting Application Areas

Trend analysis has been used to for a long time to utilize the value of historical data changes and patterns existing therein. A number of sectors have benefited from the science surrounding time series forecasting, the following are some of the areas among others.

i. Economic forecasting

This stands for predictions made about the economy. Different levels of aggregation can be used to arrive at a conclusion on economic states such as Gross Domestic Product (GDP), Inflation or Unemployment. The past data on these elements can be used to generate forecasts of what is expected at a future time period.

ii. Sales forecasting

This is an attempt to estimate future sales using past data, it allows companies to predict achievable sales revenue, plan for future growth as well as to allocate resources efficiently.

iii. Budgetary analysis

Budgetary analysis requires a set of tools to help in management of budgeting. Historical data on how different elements were required and/or consumed could be harnessed and used to predict future requirements if well modelled.

iv. Quality control

According to ISO 9000, “quality control is a part of quality management focused on fulfilling quality requirements." Quality control mainly deals with the testing of products and processes to reveal errors which will be reported to allow management decision on the product or service. Since the testing is mostly by random sampling, time series forecasting can help to find out historically where the main defects were noted to advice on where to focus the testing.

v. Inventory studies

BusinessDictionary.com defines inventory analysis as “Technique for determining the optimum level of inventory for a firm”. Analysis of how much inventory has been consumed is an important indicator of how much stock will be required in future.

vi. Workload projections

Workload analysis works to plan and predict work and skills required, this is mostly generated from past data. Historical performance adjustments can be done for workforce demand and supply changes.

vii. Census analysis

A population census is done by gathering information about the population in a country, the information is compiled, analyzed and used to assess the state of a country’s health and that of its residence. Due to the financial implications involved in conducting a census, it is conducted after a long period, mostly 10 years. In between the census, time series forecasting can be very useful to provide population projections for a year in the future before the next census.

CHAPTER 3: METHODOLOGY

3.1 Introduction

As noted by C. Kothari (2004), the formidable problem that follows the task of defining the research problem is the preparation of the design of the research project. The design of the proposed research project will be centered on the CRISP-DM Process Model as depicted in Figure 5.

3.2 Overview of Weka

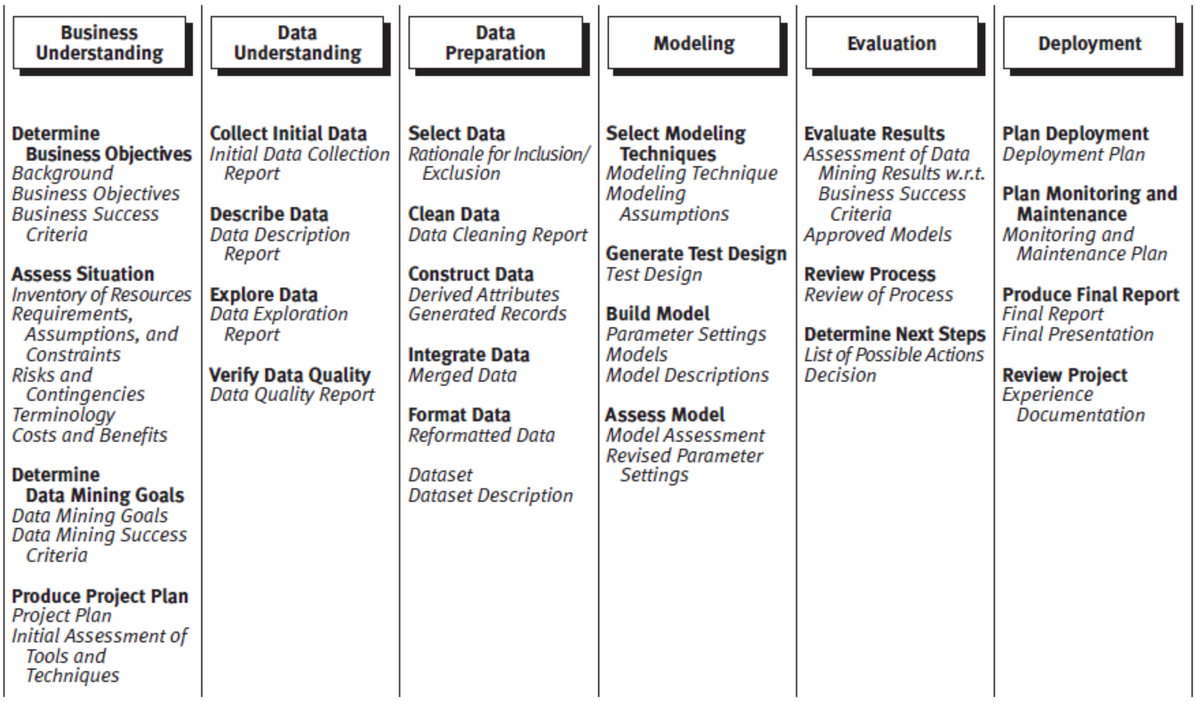
WEKA is an acronym that stands for Waikato Environment for Knowledge Analysis, its invention followed a perceived need for a single platform to allow researchers easy access to state-of the-art techniques in machine learning. The Weka project allows users to quickly try out and compare different machine learning methods on new data sets. "Weka's modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided." (Hall et. al, 2008).

The following advantages supported the decision to use Weka for this implementation

- Accessible free under the GNU General Public License.
- Portability, as it runs approximately on any modern computing platform.
- An exhaustive collection of data pre-processing and modelling techniques.
- Due to its graphical user interfaces it is Easy to use.

3.3 Application of CRISP-DM

The structure of the selected framework (CRISP-DM) was designed for easy adoption by researchers. According to the CRISP-DM Consortium, horizontally, the CRISP-DM methodology distinguishes between the reference model and the user guide. The reference model presents a quick overview of phases, tasks, and their outputs, and describes what to do in a data mining project. The user guide gives more detailed tips and hints for each phase and each task within a phase, and depicts how to carry out a data mining project. The extent of time invested in building and testing the framework was part of the motivation to select CRISP-DM as the framework for this project. The Figure below shows an overview of the CRISP-DM tasks and their outputs.



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 6: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

3.3.1 Business Understanding

3.3.1.1 Determine business objectives

The output of this project is expected to improve the methods used in generation of planning figures for refugee operations by introducing more scientific means to complement the anticipated population changes. The operational perspective was shared by the Programme Department of UNHCR and the format of the template currently used to perform the population projections was shared and saved as Appendix A. The information gathered during requirements analysis indicates that the three refugee locations in Kenya i.e Dadaab, Kakuma and Urban are required to submit at the start of every year, a three year projection of what they anticipate the refugee population would look like, this is thereafter summed up to form the projections for the country. To achieve this a number of elements below are used:

i. New arrival registration
ii. Births – calculated using the prevailing birth rate for Kenya
iii. Deaths – calculated using the prevailing mortality rate of Kenya
iv. Resettlement – expected refugee reduction due to resettlement by the office to a third country such as USA, Great Britain, Canada etc.
v. Voluntary repatriation - expected refugee returns to their country of origin

These elements are tabulated as shown in Appendix A and determine the population at the end of the year based on the population at the start of the year.

3.3.1.2 Assess situation

Apart from the birth and death rates which are calculated, the other values of elements currently used to project the refugee population at the end of the planning year are based on expected/planned activities. There exists other aspects of refugee population changes which are not included in the calculation and this is due to their lack of predictability such as refugees returning on their own without informing the office or refugees moving to a different country to seek asylum, these assumptions will hopefully be minimized by the selection of a suitable model with minimal error margin.

3.3.1.3 Determine data mining goals

Assessment of the current process of forecasting refugee population led to the development of a flow chart to help in understanding how the output is currently achieved, this is as shown in Appendix B. The understanding led to development of an AS-IS TO-BE diagram (Appendix C) mapping the comparison between the old and proposed process.

As Is Business Process

An “as is” business process defines the current state of the business process in an organization (Brandenburg, 2009). Typically the analysis goal in putting together the current state process is to clarify exactly how the business process works today. Access to the stakeholders who perform the business process is key, for this project, the Programme staff who perform the budgeting were approached for their guidance on how the process of forecasting the population is done. A walkthrough of Appendix A was conducted and it depicted the amount of work required to be done to generate the forecast, the following constraints were noted.

- Each office (Nairobi, Dadaab and Kakuma) have to perform their individual forecasts which are thereafter summed up since the portal that receives the Planning Figures requires a set of forecasts of Kenya refugees cumulatively from all refugee locations in Kenya.
- A total of 5 departments are involved in the generation of the forecasts which takes up a lot of man hours for the computations and estimates.
- Some of the estimates are based on intuition as they lack some vital facts especially after the first year forecast.

To Be Business Process

A "to be" business process defines the future state of a business process in an organization. Typically the analysis goal in putting together the future state process is to clarify how the business process will work, at some point in the future, once changes are made (Brandenburg, 2009). In the proposed procedure, the forecast will be done through learning the historical population trend of each nationality of interest, the learning will be used to create a model which will be used for the forecast. The process of creating the model will entail tuning the data mining tool to have the most optimal output with the least Mean Square Error when run on the test set data. This procedure will enable population for a number of years to be forecasted all at once as opposed to the current procedure where the forecast of one year becomes the starting figures of the following year to be taken through the same calculations and estimations.

3.3.2 Data Understanding

3.3.2.1 Collect initial data

"A data mining system has the potential of generating thousands of patterns only a small fraction of the patterns potentially generated would actually be of interest to a given user." (Han et.al, 2012). Different kinds of patterns that can be mined can be categorized as follows:

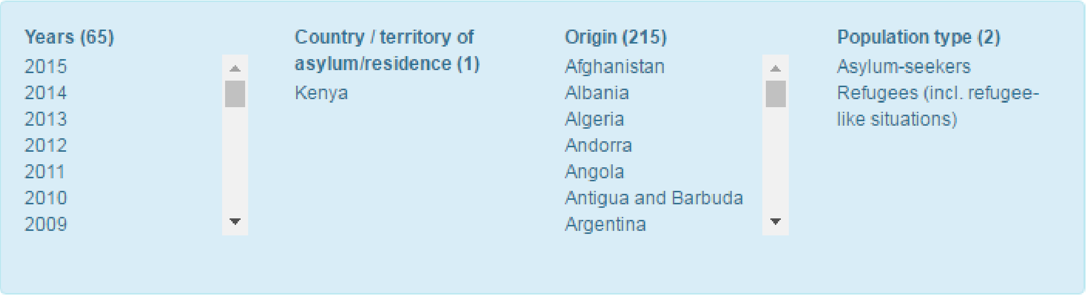
i. Class/Concept Description: Characterization and Discrimination – here, data entries can be associated with classes or concepts.
ii. Mining Frequent Patterns, Associations, and Correlations
iii. Classification and Regression for Predictive Analysis – the derived model may be represented in various forms such as classification rules, decision trees, mathematical formulae or neural networks.
iv. Cluster analysis

Data for this project will be extracted from a publicly available refugee population portal maintained by UNHCR (http://popstats.unhcr.org/) which hosts historical data of refugees and asylum seekers collected by different host countries, Kenya being one of the major asylum countries. The portal provides a facility for users to perform filters before downloading the data in CSV format, the following filters were used to extract the data required for this project.

- Reporting years: all
- Country / territory of asylum: Kenya
- Country of origin: all

- Population type: Refugees and Asylum seekers

The filter selections on the portal appear as below.



Years (65)
2015
2014
2013
2012
2011
2010
2009

Country / territory of asylum/residence (1)
Kenya

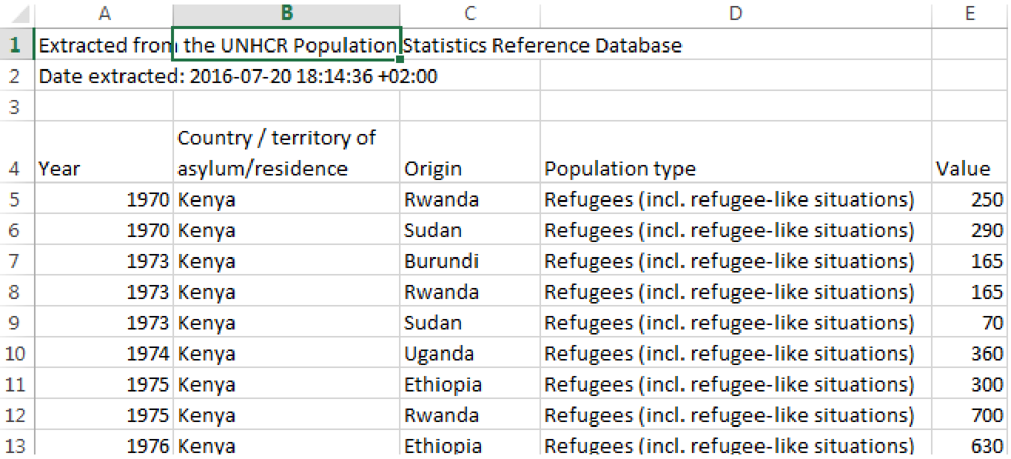
Origin (215)
Afghanistan
Albania
Algeria
Andorra
Angola
Antigua and Barbuda
Argentina

Population type (2)
Asylum-seekers
Refugees (incl. refugee-like situations)

Figure 7: Selected criteria on the UNHCR data portal (Source: http://popstats.unhcr.org/)

3.3.2.2 <u>Describe data</u>

The data collected from the portal contains basically the end year refugee population figures as reported from Kenya. A sample of the data can be found in the figure below.



	A	B	C	D	E
1	Extracted from the UNHCR Population Statistics Reference Database				
2	Date extracted: 2016-07-20 18:14:36 +02:00				
3					
4	Year	Country / territory of asylum/residence	Origin	Population type	Value
5	1970	Kenya	Rwanda	Refugees (incl. refugee-like situations)	250
6	1970	Kenya	Sudan	Refugees (incl. refugee-like situations)	290
7	1973	Kenya	Burundi	Refugees (incl. refugee-like situations)	165
8	1973	Kenya	Rwanda	Refugees (incl. refugee-like situations)	165
9	1973	Kenya	Sudan	Refugees (incl. refugee-like situations)	70
10	1974	Kenya	Uganda	Refugees (incl. refugee-like situations)	360
11	1975	Kenya	Ethiopia	Refugees (incl. refugee-like situations)	300
12	1975	Kenya	Rwanda	Refugees (incl. refugee-like situations)	700
13	1976	Kenya	Ethiopia	Refugees (incl. refugee-like situations)	630

Figure 8: Sample data extracted from the UNHCR refugee statistics public portal (http://popstats.unhcr.org/)

From the surface, the data contains population figures of refugees and asylum seekers in Kenya reported from as far back as 1970. The current procedure of extrapolating the population planning figures focuses on main refugee nationalities and groups the minor nationalities into an “Other” category. The main countries of origin by June 2016 were: Burundi, Dem. Republic of the Congo, Eritrea, Ethiopia, Rwanda, Somalia, South Sudan, Sudan and Uganda. A total of 871 rows of data dating as far back as 1970 were extracted.

3.3.2.3 Explore data

This task was used to try and addresses data mining questions using data transformation and visualization to help understand the data. These include the relationships or similarities between different nationalities etc. The proponents of this model state that this task may directly address the data mining goals or help to refine the data description and quality reports.

The extracted data was placed in a Microsoft Excel document and a Pivot Table and Pivot Chart used to re-arrange and explore the data. At the onset it was clear that data for Asylum Seekers was only available from the year 2000 in all the nationalities, the figure below depicts the initial findings from the exploration of extracted data.

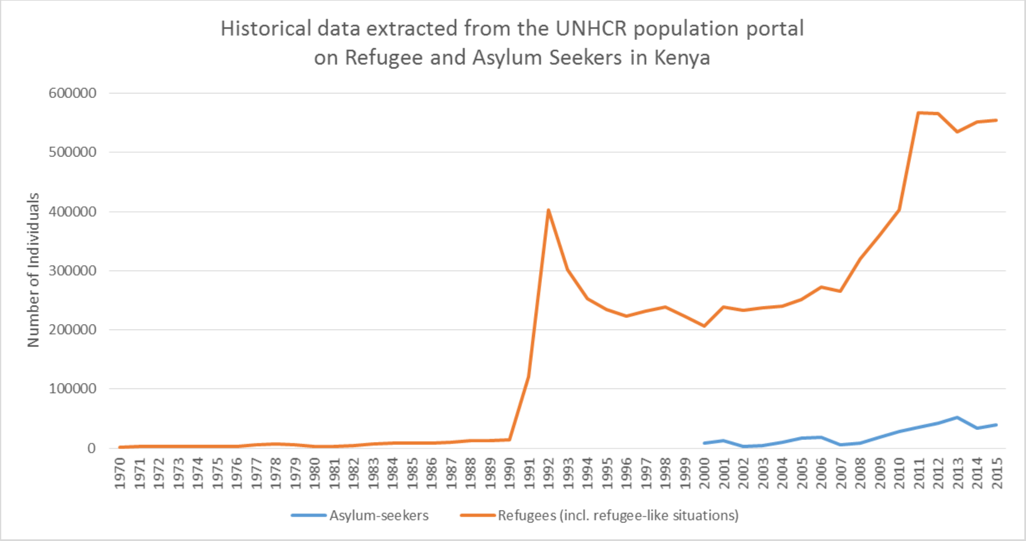


Figure 9: Visualization of exploration exercise on extracted data

It is additionally clear that there is an upward trend of the population of the refugee and asylum seeker population in Kenya, additionally there exists two specific spikes in population in 1990 to 1992 and 2010 to 2011.

3.3.2.4 Verify data quality

This step involved the examination of the quality of data for completeness, to ensure that it covers all the cases required. Since the data was not collected and entered manually, there were no visible errors or data type issues, the data was therefore ready for the next CRISP-DM step of “Data Preparation”.

3.3.3 Data Preparation

3.3.3.1 Select data

In a September 2004 article available in www.unhcr.org, Goldstein-Rodriguez outlined the launch of a new refugee registration system called Profile Global Registration System (proGres) which according to the writer was replacing dozens of databases that were not necessarily compatible. It is noted that "proGres" is still in use in 2016 having evolved through a number of versions and improvements. The year 2004 was therefore selected as the cutoff date for selection of the data to be used for analysis to create a model. In addition UNHCR introduced biometric registration shortly after thereby increasing credibility in data quality as individuals could be registered only once due to the uniqueness of biometric identity.

3.3.3.2 Clean data

Following the data exploration phase above as well as data quality verification and selection, it was noted that there were some records which would not be useful for the intended analysis, the following are the categories that will be eliminated:

- Nationalities which were indicated as "Various/Unknown".
- Data that was recorded before the year 2004 as noted in the Data Selection phase.
- Records containing an asterix (*) in the values field. The figures in these fields were redacted as per comments from the portal, the missing figures would however not cause a big effect on the resulting model since they only affected a few minority nationalities in the records of the year 2014 and 2015.

After the data cleanup task, the number of records that remained were 517 out of the 871 that were extracted initially from the UNHCR public statistics portal.

3.3.3.3 Construct data

This task includes constructive data preparation operations such as the production of derived attributes or entire new records, or transformed values for existing attributes. The data construction tasks involved in this project are:

- Combination of data from the minority nationalities into the "Other" category.
- Estimation of South Sudanese and Sudanese data from 2004 to 2011.
- Summation of the refugee figures to the asylum seekers figures to get a total figure per nationality.

  Total population per nationality = Refugee population + Asylum seeker population

Estimation of Sudan and South Sudan data

South Sudan separated from the Republic of Sudan in the aftermath of the January 2011 referendum that separated the two (Belloni, 2011). Statistically, the two countries were reported separately from the year 2012. To estimate how the two populations we represented from 2004 to 2011, the 2012 data was used which revealed that South Sudan consisted of 85% of the total data of Sudan and South Sudan, this ratio was therefore applied to the missing data as the closest possible estimation to allow creation of a model for the two countries.

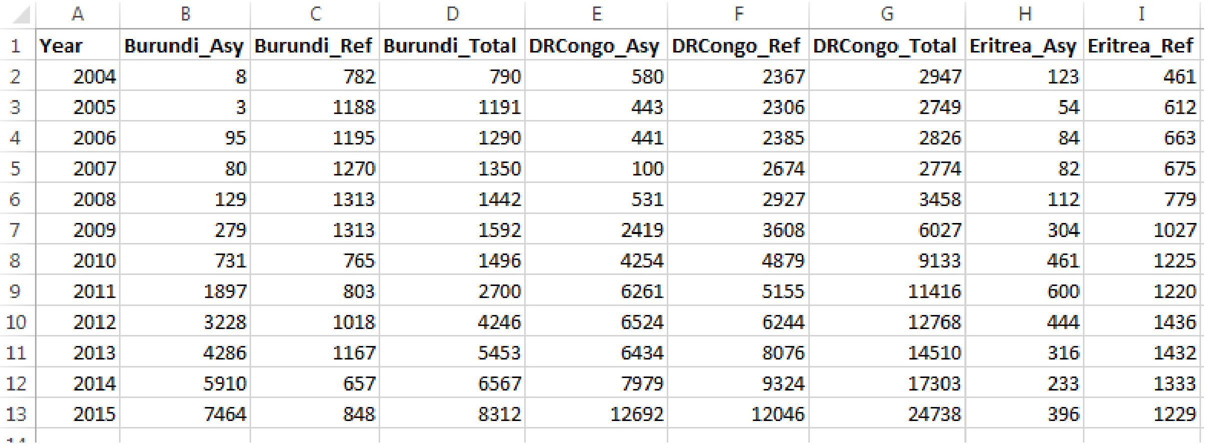
3.3.3.4 Format data

Very minor formatting tasks were required for the data. Weka requires attribute names which do not have spaces in between, the following attributes were therefore changed to comply with the requirements of the modelling platform.

- Dem. Rep. of the Congo → DRCongo
- South Sudan → S.Sudan

In addition, since each nationality is split into two population groups 'refugee' and 'asylum-seeker', the suffixes '_Ref' and '_Asy' were used to form two unique attributes per nationality, the totals of each nationality were represented with the suffix '_Total'. For example Somalia was represented by the following attributes: Somalia_Ref, Somalia_Asy and Somalia_Total.

The resultant data was stored in comma separated values (CSV) file format which is compatible with Weka, the data appears as the below sample:



	A	B	C	D	E	F	G	H	I
1	Year	Burundi_Asy	Burundi_Ref	Burundi_Total	DRCongo_Asy	DRCongo_Ref	DRCongo_Total	Eritrea_Asy	Eritrea_Ref
2	2004	8	782	790	580	2367	2947	123	461
3	2005	3	1188	1191	443	2306	2749	54	612
4	2006	95	1195	1290	441	2385	2826	84	663
5	2007	80	1270	1350	100	2674	2774	82	675
6	2008	129	1313	1442	531	2927	3458	112	779
7	2009	279	1313	1592	2419	3608	6027	304	1027
8	2010	731	765	1496	4254	4879	9133	461	1225
9	2011	1897	803	2700	6261	5155	11416	600	1220
10	2012	3228	1018	4246	6524	6244	12768	444	1436
11	2013	4286	1167	5453	6434	8076	14510	316	1432
12	2014	5910	657	6567	7979	9324	17303	233	1333
13	2015	7464	848	8312	12692	12046	24738	396	1229

Figure 10: Excerpt of the data after data preparation phase

3.3.4 Modelling

3.3.4.1 Select modelling technique

When the outcome, or class, is numeric, and all the attributes are numeric, regression is a natural technique to consider (Witten et. al., 2011). The version of Weka selected was version 3.8.0 which was the latest version by the time of execution of this project. This version contains an installable forecasting module introduced in version 3.7.x which is the main module used in this project for Time Series Forecasting. Weka supports many methods for predicting time series as function-based learning, these include: Linear regression, Multilayer perceptron neural network, SMOreg-support vector machine for regression. The functions will be compared for suitability to create a model and eventually only one will be selected. Brownlee (2016) used the Root Mean Squared Error (RMSE) to compare between different regression machine learning algorithms in Weka, he used RMSE to rank them. For this algorithm selection process we will use the Mean Absolute Percentage Error (MAPE) in addition to RMSE. The data from the total Burundi population from 2004 to 2015 as shown in figure above was used for the evaluation.

Linear regression

"It is good idea to evaluate linear regression on a data mining problem before moving onto more complex algorithms in case it performs well. It is a very simple regression algorithm, fast to train and can have great performance if the output variable for the data is a linear combination of the inputs." (Brownlee, 2016). The following readings were achieved from the evaluation.

• MAPE	0
• RMSE	0.0006

Multilayer perceptron

"The Multi-Layer Perceptron algorithms also known as artificial neural networks, supports both regression and classification problems. It is an algorithm inspired by a model of biological neural networks in the brain where small processing units called neurons are organized into layers that if configured well are capable of approximating any function. Neural networks are a complex algorithm to use for predictive modeling because there are so many configuration parameters that can only be tuned effectively through intuition and a lot of trial and error." (Brownlee, 2016). The following readings were achieved from the evaluation.

MAPE	3.255
RMSE	107.2685

Support Vector Machine for regression (SMOreg)

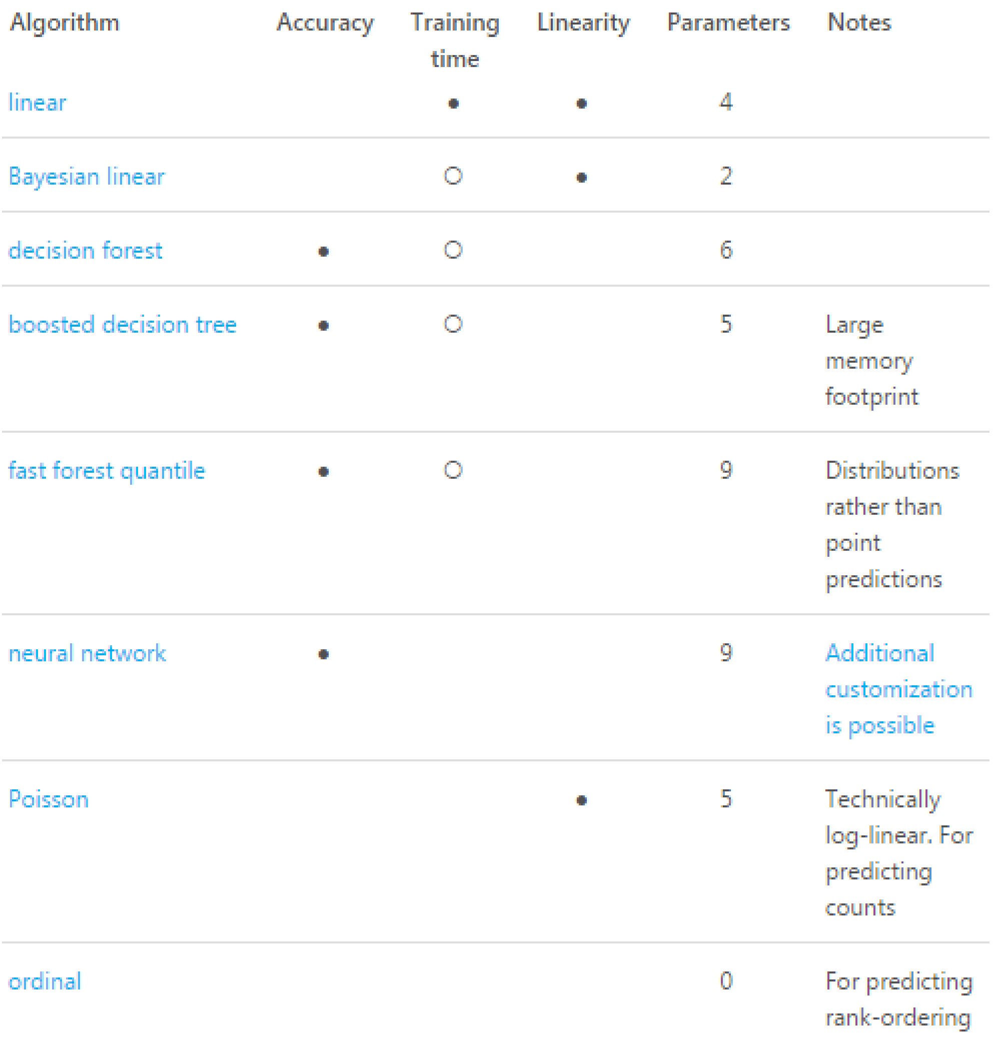
Support Vector Machines were developed for binary classification problems, although extensions to the technique have been made to support multi-class classification and regression problems. It works by using an optimization process that only considers those data instances in the training dataset that are closest to the line with the minimum cost. These instances are called support vectors, hence the name of the technique. The following readings were achieved from the evaluation.

MAPE	10.4026
RMSE	349.14

Selected Algorithm

Linear regression appeared to perform well in the evaluation however the MAPE of 0% was not realistic for the dataset processed. SMOReg performed well however had a MAPE that was higher than 10% which was considered to be quite high for the intended forecasting. The Multilayer Perceptron was quite impressive in the evaluation, it produced the model with a MAPE of 3% which is quite acceptable for use in developing the forecasting model.

Microsoft Azure is a big player in the space of Machine Learning and Big Data, in the online article by Rohrer (2016), the writer provided some tips on how to choose Machine Learning algorithms and summarized it in the table below.



Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
linear		●	●	4	
Bayesian linear		○	●	2	
decision forest	●	○		6	
boosted decision tree	●	○		5	Large memory footprint
fast forest quantile	●	○		9	Distributions rather than point predictions
neural network	●			9	Additional customization is possible
Poisson			●	5	Technically log-linear. For predicting counts
ordinal				0	For predicting rank-ordering

Figure 11: Tips on selecting Machine Learning Algorithms (Source: Azure.microsoft.com)

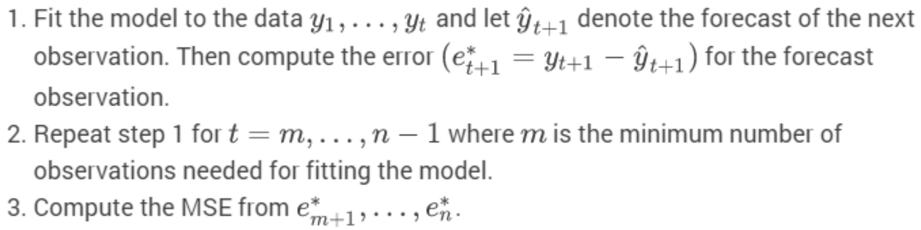
The table above shows neural networks featuring well in accuracy as well as additional customization which is good for fine-tuning the output. Following the findings of the evaluation as well as literature review, Multilayer Perceptron was selected to create the model for this project.

3.3.4.2 <u>Generate test design</u>

A testing sequence independent of the training data set and distributed according to the same probability distribution is used to assess the quality of the learning process Bontempi (2013), unfortunately an additional set of input/output observations is rarely available. A Holdout method (also known as test sample estimation) is used to partition the data into two mutually exclusive subsets, the training data set and the holdout or test set. "In most real applications, only a limited amount of data is available." (Arlot and Celisse, 2009), which leads to the idea of splitting the data: Part of data (the training sample) is used for training the algorithm, and the remaining data (the validation sample) are used for evaluating the performance of the algorithm. The validation sample can play the role of “new data” before the model is built, a procedure should be generated or a mechanism to test the model’s quality and validity.

Cross-validation for time series forecasting

Various methods of estimating a model’s accuracy exist such as the K-fold cross-validation, leave one out cross-validation, repeated learning testing among other variants. Majority of these are suited to classification models and not time series forecasting. When the data are not independent cross-validation becomes more difficult as leaving out an observation does not remove all the associated information due to the correlations with other observations (Hyndman, 2010). For time series forecasting, a cross-validation statistic is obtained as per the excerpt below:

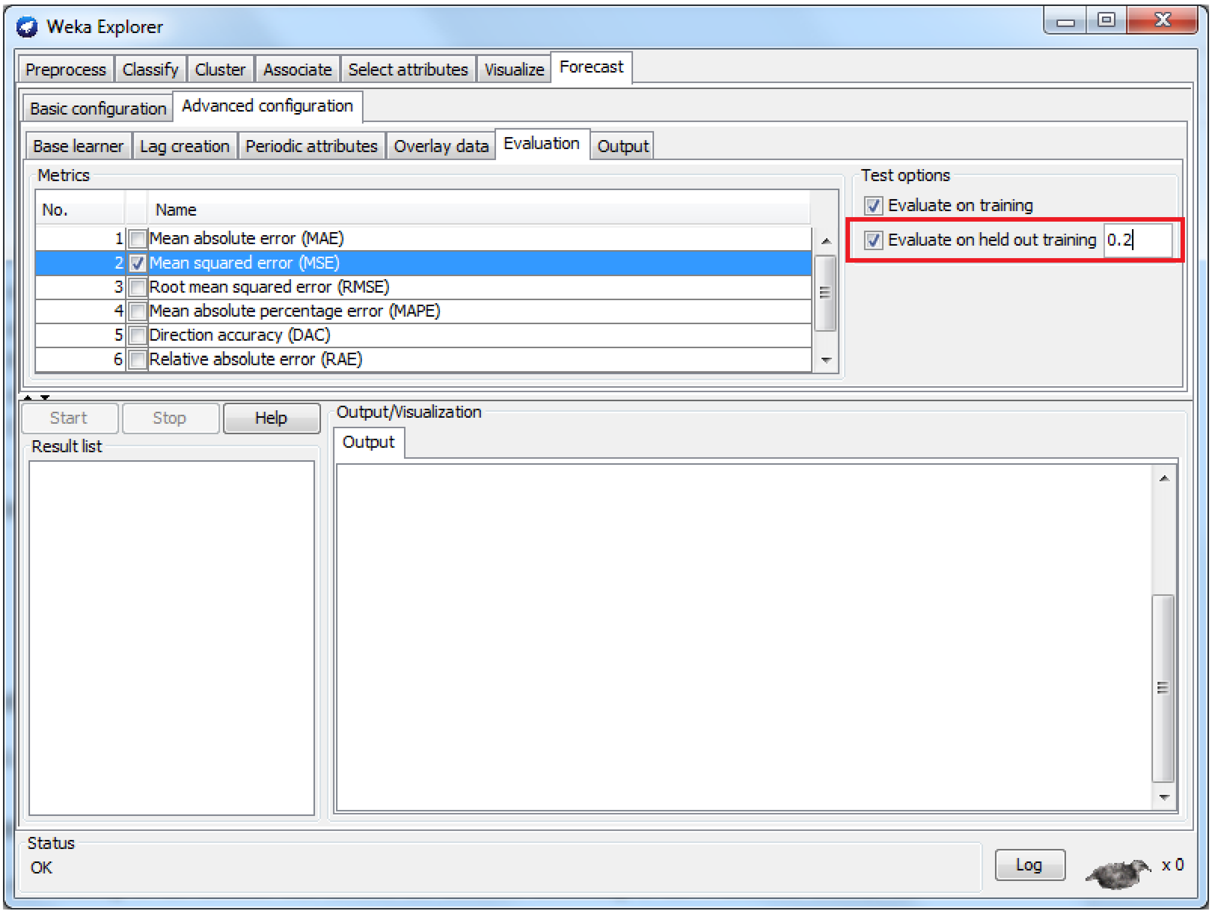


1. Fit the model to the data $y_1, \ldots, y_t$ and let $\hat{y}_{t+1}$ denote the forecast of the next observation. Then compute the error $(e^*_{t+1} = y_{t+1} - \hat{y}_{t+1})$ for the forecast observation.
2. Repeat step 1 for $t = m, \ldots, n-1$ where m is the minimum number of observations needed for fitting the model.
3. Compute the MSE from $e^*_{m+1}, \ldots, e^*_n$.

Figure 12: Cross-validation for time series (Source: http://robjhyndman.com/*)*

Holdout validation for time series forecasting

Generally, the larger the training sample, the better the model (Witten et.al, 2011). The researchers continue to add that the real problem occurs when there is not a vast supply of data available. The dataset used for this project contains a 12 year refugee data from the year 2004 to 2015. The minimum number of observations needed for fitting the model will be taken as 80% of the entire dataset therefore 20% will be used as the test set to evaluate the model. Weka provides in built tools to from the Graphical User Interface to perform majority of functions that are required in implementing a data mining project, the holdout validation will be performed using the tools available under the “Advanced Configuration” tab of the Forecast module in Weka as shown in the figure below.



Metrics

No.	Name
1	□ Mean absolute error (MAE)
2	☑ Mean squared error (MSE)
3	□ Root mean squared error (RMSE)
4	□ Mean absolute percentage error (MAPE)
5	□ Direction accuracy (DAC)
6	□ Relative absolute error (RAE)

Test options

☑ Evaluate on training

☑ Evaluate on held out training 0.2

Status

OK

Figure 13: Configuration of test set data for model evaluation in Weka

The option in the red boundary in the Figure above represents the 20% of the dataset that will be used for testing. Weka additionally has the option of selecting the number of records to be used for testing, this can be done by entering a whole number in the field highlighted in the image above.

3.3.4.3 Build model

The business understanding phase as well as data exploration in the data understanding phase revealed the distinct nature of the population trends for different nationalities, the total population therefore cannot be forecasted all at once, it will done based on the sum of forecasts from different nationalities. In the current procedure, refugee arrivals are estimated based on the prevailing conditions in the different countries of origin; refugee resettlement to a third country is also based on individual country estimates. Mortality rate as well as birth rate are calculated based on the prevailing rates in Kenya, these will therefore affect the end year populations by a standard ratio. Below is a summary of the variables currently used to project refugee populations.

Dependent on Nationality		**Not Dependent on Nationality**	
Population increase	**Population decrease**	**Population increase**	**Population decrease**
New arrivals	Resettlement Repatriation Local integration	New birth	Death

Table 1: Variables used to calculate population forecast and their relationship to the refugee Nationality

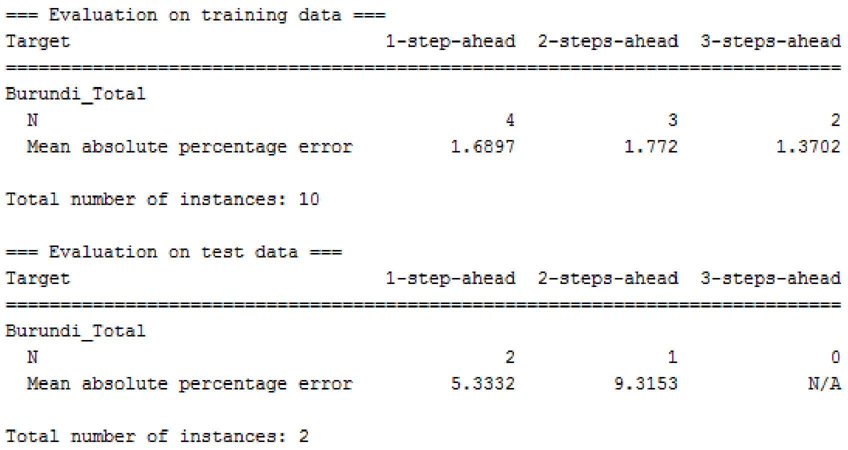
The following sections will provide details and configuration of the models generated from Weka per nationality, minority nationalities are consolidated under “Other nationalities”. Evaluation was done on the training data as well as test data to avoid overfitting, the forecast was done for three years. The Mean Absolute Percentage Error (MAPE) was used in the process of tuning the model, the least MAPE on both training and test data was considered most optimal for the model.

Burundi

The model for Burundi was built using ‘Burundi_Total’ attribute of the dataset, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 3, Maximum 6

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                               1-step-ahead  2-steps-ahead  3-steps-ahead
================================================================================
Burundi_Total
  N                                             4              3              2
  Mean absolute percentage error           1.6897          1.772         1.3702

Total number of instances: 10

=== Evaluation on test data ===
Target                               1-step-ahead  2-steps-ahead  3-steps-ahead
================================================================================
Burundi_Total
  N                                             2              1              0
  Mean absolute percentage error           5.3332         9.3153            N/A

Total number of instances: 2
```

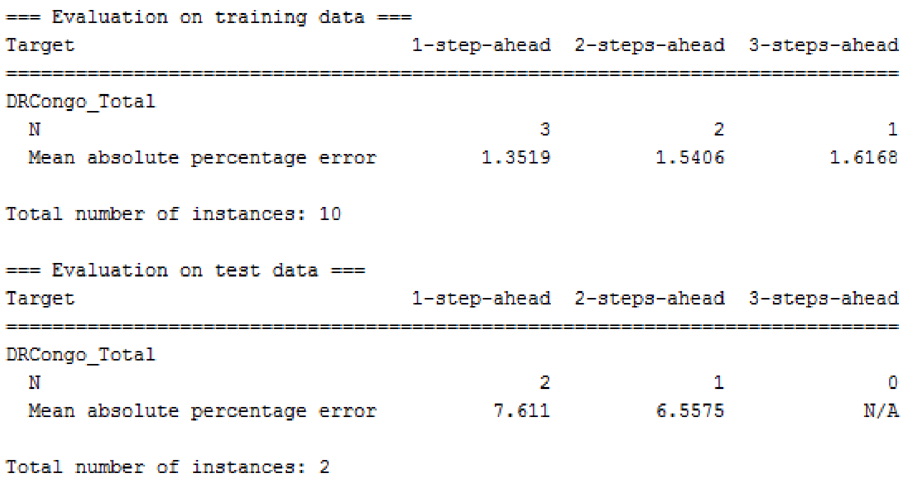
Figure 14: Output of Burundi model evaluation

Congo DR

The model for Congo Democratic Republic was built using ‘DRCongo_Total’ attribute of the dataset, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 6, Maximum 7

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
DRCongo_Total
  N                                            3              2              1
  Mean absolute percentage error          1.3519         1.5406         1.6168

Total number of instances: 10

=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
DRCongo_Total
  N                                            2              1              0
  Mean absolute percentage error           7.611         6.5575            N/A

Total number of instances: 2
```

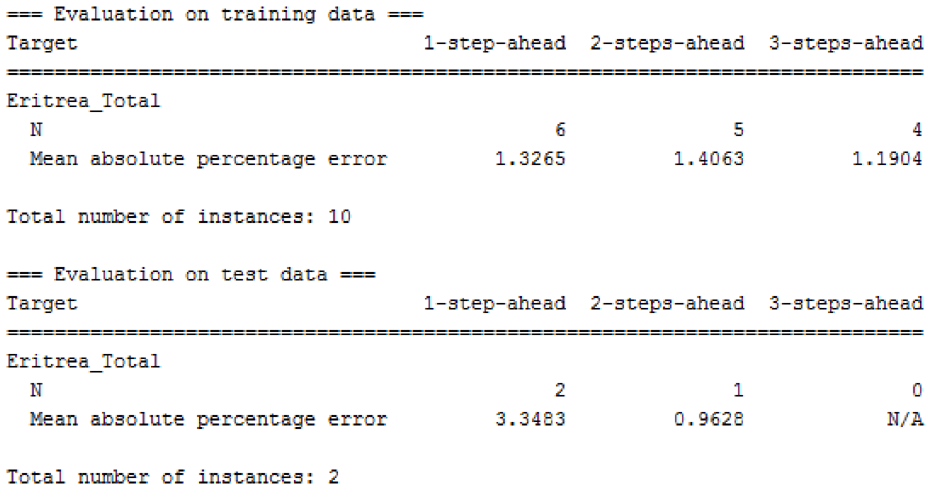
Figure 15: Output of Congo DR model evaluation

Eritrea

The model for Eritrea was built using ‘Eritrea_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 3, Maximum 4

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                                 1-step-ahead  2-steps-ahead  3-steps-ahead
=================================================================================
Eritrea_Total
  N                                               6              5              4
  Mean absolute percentage error             1.3265         1.4063         1.1904

Total number of instances: 10

=== Evaluation on test data ===
Target                                 1-step-ahead  2-steps-ahead  3-steps-ahead
=================================================================================
Eritrea_Total
  N                                               2              1              0
  Mean absolute percentage error             3.3483         0.9628            N/A

Total number of instances: 2
```

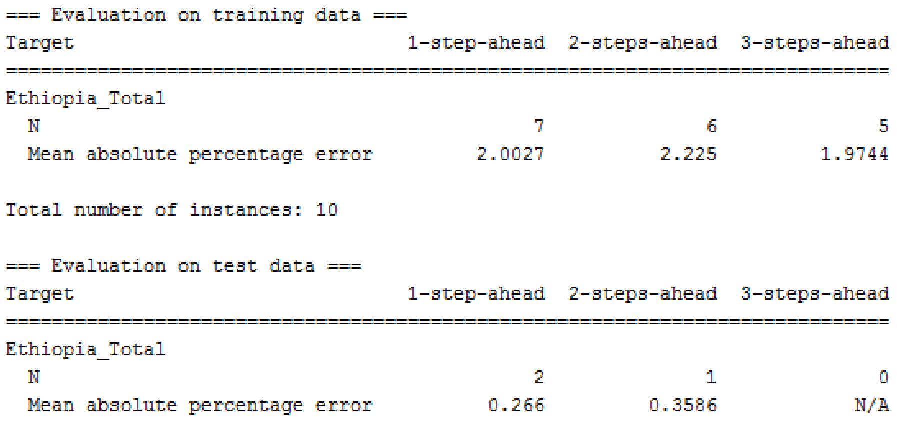
Figure 16: Output of Eritrea model evaluation

Ethiopia

The model for Ethiopia was built using ‘Ethiopia_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 3, Maximum 3

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                                 1-step-ahead  2-steps-ahead  3-steps-ahead
=================================================================================
Ethiopia_Total
  N                                               7              6              5
  Mean absolute percentage error             2.0027          2.225         1.9744

Total number of instances: 10

=== Evaluation on test data ===
Target                                 1-step-ahead  2-steps-ahead  3-steps-ahead
=================================================================================
Ethiopia_Total
  N                                               2              1              0
  Mean absolute percentage error              0.266         0.3586            N/A
```

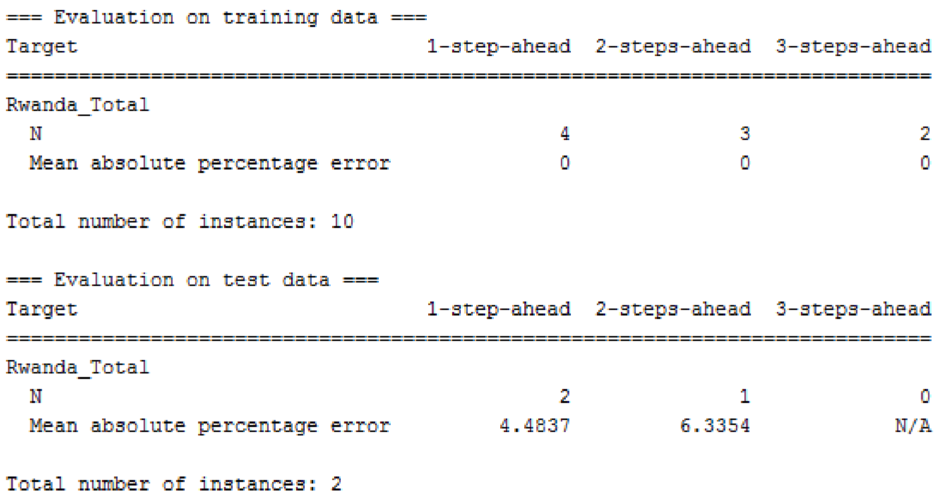
Figure 17: Output of Ethiopia model evaluation

Rwanda

The model for Rwanda was built using ‘Rwanda_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 1, Maximum 6

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                          1-step-ahead  2-steps-ahead  3-steps-ahead
==========================================================================
Rwanda_Total
  N                                        4              3              2
  Mean absolute percentage error           0              0              0

Total number of instances: 10

=== Evaluation on test data ===
Target                          1-step-ahead  2-steps-ahead  3-steps-ahead
==========================================================================
Rwanda_Total
  N                                        2              1              0
  Mean absolute percentage error      4.4837         6.3354            N/A

Total number of instances: 2
```

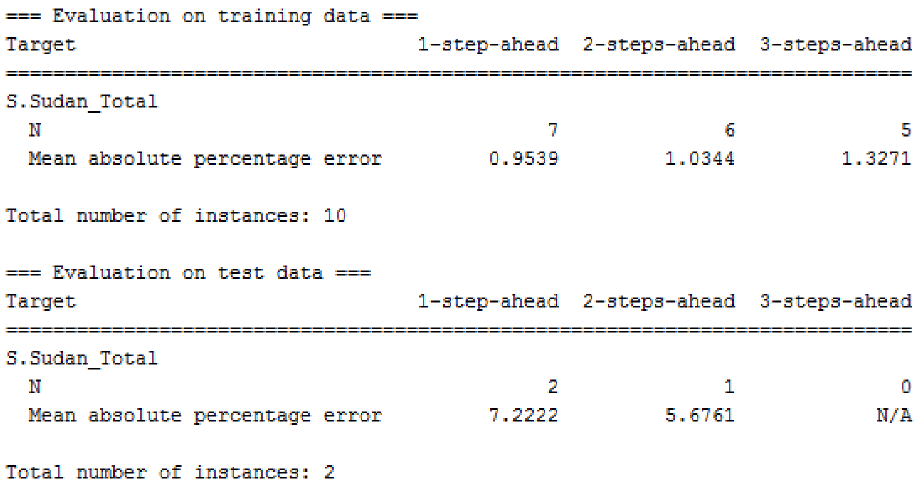
Figure 18: Output of Rwanda model evaluation

South Sudan

The model for South Sudan was built using ‘S.Sudan_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 1, Maximum 3

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                          1-step-ahead  2-steps-ahead  3-steps-ahead
==========================================================================
S.Sudan_Total
  N                                        7              6              5
  Mean absolute percentage error      0.9539         1.0344         1.3271

Total number of instances: 10

=== Evaluation on test data ===
Target                          1-step-ahead  2-steps-ahead  3-steps-ahead
==========================================================================
S.Sudan_Total
  N                                        2              1              0
  Mean absolute percentage error      7.2222         5.6761            N/A

Total number of instances: 2
```

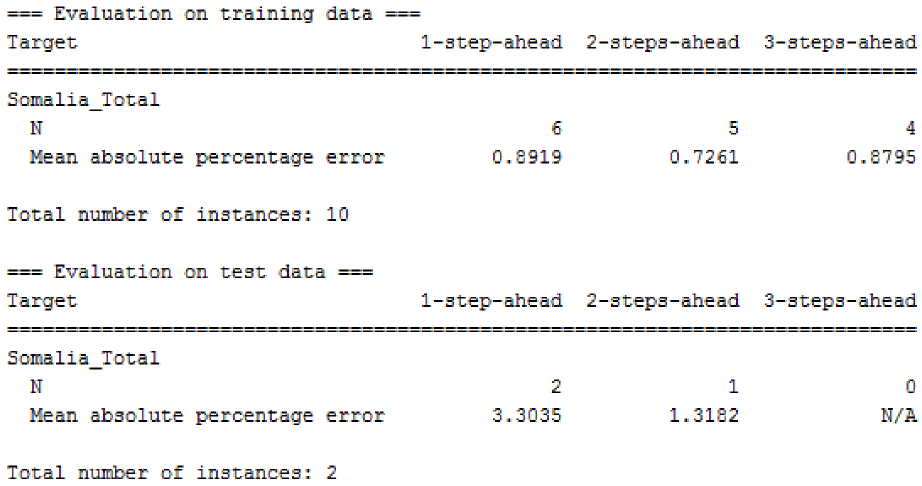
Figure 19: Output of South Sudan model evaluation

Somalia

The model for Somalia was built using ‘Somalia_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 1, Maximum 4

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                           1-step-ahead  2-steps-ahead  3-steps-ahead
===========================================================================
Somalia_Total
  N                                         6              5              4
  Mean absolute percentage error       0.8919         0.7261         0.8795

Total number of instances: 10

=== Evaluation on test data ===
Target                           1-step-ahead  2-steps-ahead  3-steps-ahead
===========================================================================
Somalia_Total
  N                                         2              1              0
  Mean absolute percentage error       3.3035         1.3182            N/A

Total number of instances: 2
```

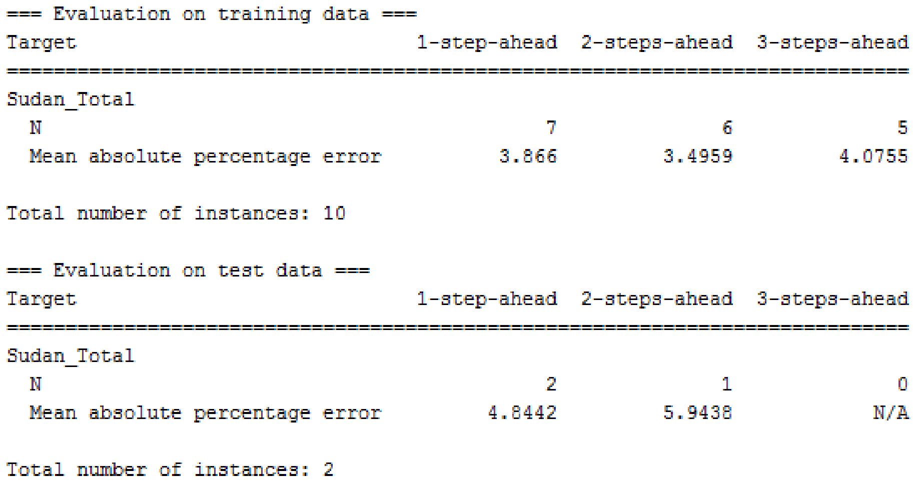
Figure 20: Output of Somalia model evaluation

Sudan

The model for Sudan was built using 'Sudan_Total' attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 3, Maximum 3

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                           1-step-ahead  2-steps-ahead  3-steps-ahead
===========================================================================
Sudan_Total
  N                                         7              6              5
  Mean absolute percentage error        3.866         3.4959         4.0755

Total number of instances: 10

=== Evaluation on test data ===
Target                           1-step-ahead  2-steps-ahead  3-steps-ahead
===========================================================================
Sudan_Total
  N                                         2              1              0
  Mean absolute percentage error       4.8442         5.9438            N/A

Total number of instances: 2
```

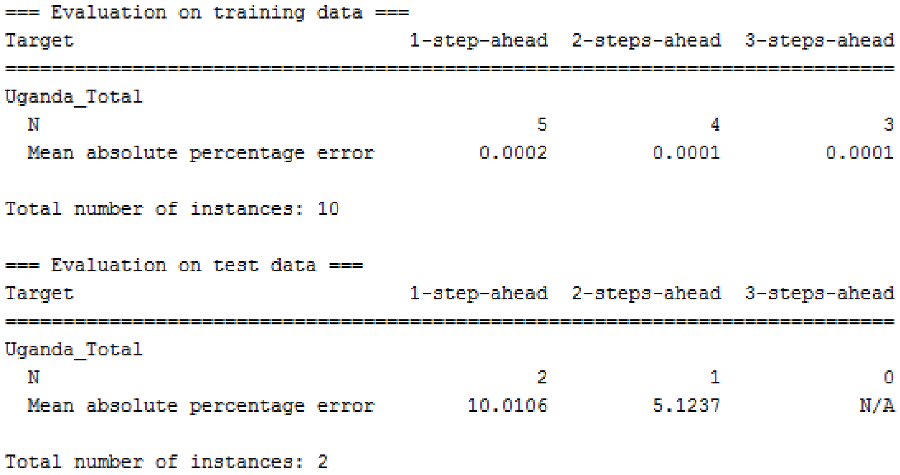
Figure 21: Output of Sudan model evaluation

Uganda

The model for Uganda was built using 'Uganda_Total' attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 1, Maximum 7

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
Uganda_Total
  N                                            5              4              3
  Mean absolute percentage error          0.0002         0.0001         0.0001

Total number of instances: 10

=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
Uganda_Total
  N                                            2              1              0
  Mean absolute percentage error         10.0106         5.1237            N/A

Total number of instances: 2
```

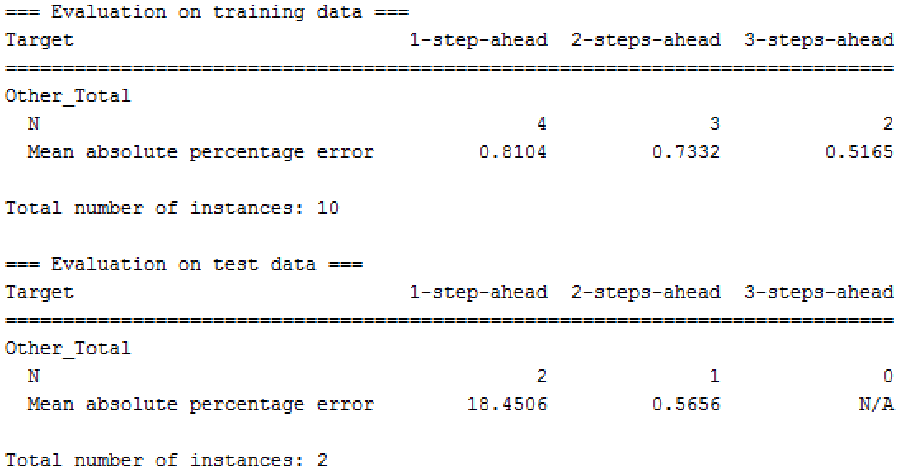
Figure 22: Output of Uganda model evaluation

Other

The model for ‘Other’ nationalities combined was built using ‘Other_Total’ attribute, the following parameters were set to reach the most optimal model.

Custom Lag: Minimum 4, Maximum 6

The output below was achieved from the above parameter settings which provided the least MAPE, this setting was used to create the model.



```
=== Evaluation on training data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
Other_Total
  N                                            4              3              2
  Mean absolute percentage error          0.8104         0.7332         0.5165

Total number of instances: 10

=== Evaluation on test data ===
Target                              1-step-ahead  2-steps-ahead  3-steps-ahead
==============================================================================
Other_Total
  N                                            2              1              0
  Mean absolute percentage error         18.4506         0.5656            N/A

Total number of instances: 2
```

Figure 23: Output of ‘Other nationalities’ model evaluation

3.3.4.4 Assess model

The process of model building iteratively involved assessment. Only the runs which satisfied the evaluation criteria of least MAPE was selected to create the final model for each nationality.

3.4 Dealing with uncertainty

"The uncertainty of population forecasts depends on the uncertainty of forecasts of fertility, mortality, and international migration. The uncertainty of forecasts of these three demographic components depends on the question in what way their future development may be different from the past or different in another way than expected." (De Beer, 2000). Uncertainty exists in forecasting because it is uncertain whether the factors affecting population will remain stable in the future.

The process of generating prediction models in Weka manages uncertainty to some extent by revealing the existence of forecasting errors calculated through model evaluation, this therefore allows the consumers of the resulting data to be aware of a range of values that can be considered acceptable in each context. This appeared to work well for nationalities which had a fairly consistent rate of population change such as Burundi, Eritrea, Ethiopia and Uganda. Two countries revealed the need to handle uncertainty differently, the details are as below.

i. Somalia – in November 2014 the Government of Kenya, UNHCR and the Government of Somalia signed a Tripartite Agreement on the voluntary return of Somali refugees back to their country of origin, this will effectively see the reduction in the population of Somali refugees in Kenya. The number targeted for return in every year cannot be learned by the forecasting model rather the agencies managing the planning of the exercise would have this information.
ii. South Sudan – in December 2012, violence broke out in South Sudan leading to millions been displaced from their homes. This led to a refugee influx in the neighboring countries of Ethiopia, Uganda and Kenya. The recurrence of such an event can only be predicted through following information from the media regarding the political and security situation as well as monitoring trends in refugee arrivals from that country.

Though addressing this additional information does not entirely eliminate the forecast uncertainty, there is need to factor it in as it reduces considerably errors due to uncertain events. Tayman (2009) outlines certain challenges faced by forecasters and outlines three common responses to forecast uncertainty as follows.

a) Alternative scenarios – this provides a range of projections in the form of an upper limit and a lower limit and gives users choices, it however fails to capture real world fluctuations.

b) Statistical probability intervals - this provides probability statements to measure forecast uncertainty.

c) Involving experts – this could use the same models and assumptions as an official forecast. The researcher however argues that it could be influenced by dominant personalities.

Selected response to uncertainty

The issue of uncertainty in forecasts has been of interest to many researchers in that area; Keilman (2001) weighs in by stating that judgmental methods can be used to correct or constrain broad prediction intervals. Expert judgement is also used when expected values and corresponding prediction intervals are hard to obtain by formal methods. A combination of expert judgement with time-series models was employed by Lutz et.al (2000) and generated some interesting results. This was adopted in this project as the means to handle uncertainty. The expert input will be sought from the three main domains which affect refugee population in Kenya outlined below.

i. Resettlement - the relocation of refugees to a third country such as USA, Canada, Australia etc as a means of burden sharing with the host country (Kenya). The resettlement programme projects roughly how many cases can be relocated per year, this can be coded into the factors to affect the population.

ii. Voluntary Repatriation – this is the voluntary movement of refugees back to their country of origin, this is assisted by the Government of Kenya and humanitarian actors. The Government announced in the year 2016 an increased number of Somali refugees will be repatriated, this will lead to a sharp reduction of the Somali population which could not be learnt by the trained model.

iii. Population Influx – this is a sharp increase in the arrival of refugees due to deteriorating conditions in the country of origin.

CHAPTER 4 : RESULTS AND DISCUSSION

4.1 <u>Evaluation of Results</u>

This project aimed to use historical data and time series forecasting to generate models which can be used to predict future refugee populations in in Kenya. The table below shows the predictions for the year 2016 performed by the generated models for each nationality compared to the actual population statistics reported by September 2016.

The table contains the following elements for each nationality:

- September 2016 statistics – this contains the population figures reported in Kenya by UNHCR
- Prediction by current procedure - this is the prediction generated at the start of the year 2016 using the current procedure using the template in Appendix A and the procedure explained in Appendix B.
- Prediction by prototype - this is the output from the developed prototype consisting of the model predictions coupled with the input from the experts. It is calculated as follows.

$$\begin{array}{ccccccc} \textit{Model} & \_ & \textit{Voluntary} & \_ & \textit{Resettlement} & & \textit{New Arrival Registration} \\ \textit{Prediction} & & \textit{Repatriation} & & \textit{projection} & + & \textit{(Population influx)} \end{array}$$

- The percentage difference in the two last columns is calculated based on the difference between the prediction approaches and the most recent statistics reported in September 2016.

Country of Origin	September 2016 Statistics	Prediction by Current Procedure	Prediction by Prototype	Current procedure % Difference	Proposed Prototype % Difference
Burundi	7,964	10,131	9,584	-27%	-20%
Dem. Rep of Congo	27,555	27,010	29,849	2%	-8%
Eritrea	1,551	1,570	1,676	-1%	-8%
Ethiopia	26,754	28,738	29,020	-7%	-8%
Rwanda	1,559	1,510	1,575	3%	-1%
South Sudan	91,212	104,467	93,883	-15%	-3%
Somalia	332,725	373,187	372,836	-12%	-12%
Sudan	9,901	11,391	10,916	-15%	-10%
Uganda	2,012	2,414	2,376	-20%	-18%
Other	298	245	312	18%	-5%

Table 2: Comparison of predicted population by current procedure and proposed prototype against actual refugee population in Kenya reported at the end of September 2016.

The expert input data was obtained from projections of Voluntary Repatriation, Resettlement and New arrival registration carried out by UNHCR at the start of 2016. From the findings it is clear that the models performed quite well in comparison to the real world data. That is an indication that the training process was well configured and the input historical data was helpful in providing a trend which could be understood by machine learning techniques.

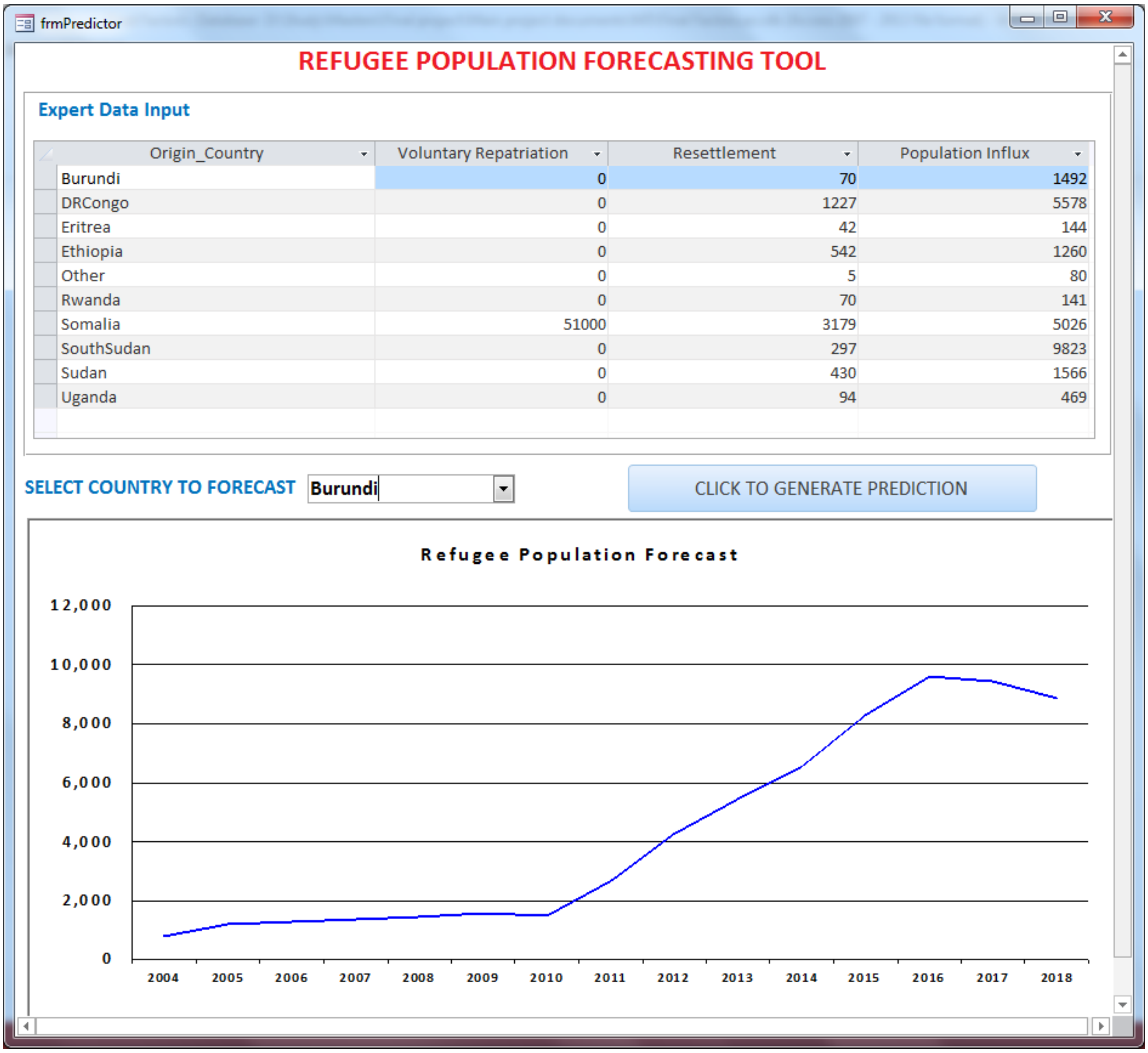
The generation of models was however not standard across all nationalities, it was found that models for nationalities which had more consistent population changes such as Congo Democratic Republic, Burundi and Ethiopia; the prediction by the models was fairly consistent with the expected population at the end of the year. A country such as Somalia which are a target group for a voluntary repatriation programme (leading to reduced numbers by end year) or South Sudan which has had a long standing political instability (leading to possibility of increased numbers by end year) could not be predicted accurately without the introduction of input from experts outlined by Tayman (2009). The experts basically estimate the population change due to activities in their domains for example, the experts in the Resettlement section would use their targeted number of individuals to be Resettled and the likelihood of achieving the targets based on the resources allocated, experts monitoring refugee arrivals to Kenya could use the daily or monthly trends as well as mass media information on the possibility of civil unrest etc. in the affected countries.

Process review

The research methods implemented in this project will allow for future researchers to use the data and informational resources gathered. The data source identified is a publicly available refugee population statistics repository which is updated by UNHCR at the end of every year for every country in the world hosting refugees and asylum seekers allowing the learning models to be created with every new data annually, the more the training data provided the better the evaluation and learning. This additionally would help researchers all over the world to adopt and localize this research to their countries and regions. Weka software used for creating the models is available for researchers without any licensing.

4.2 Implementation of uncertainty management

The forecasting tool was designed to include a facility for experts to provide their expected figures. The figures below show the predictions generated for selected nationalities incorporating the input from experts.



frmPredictor

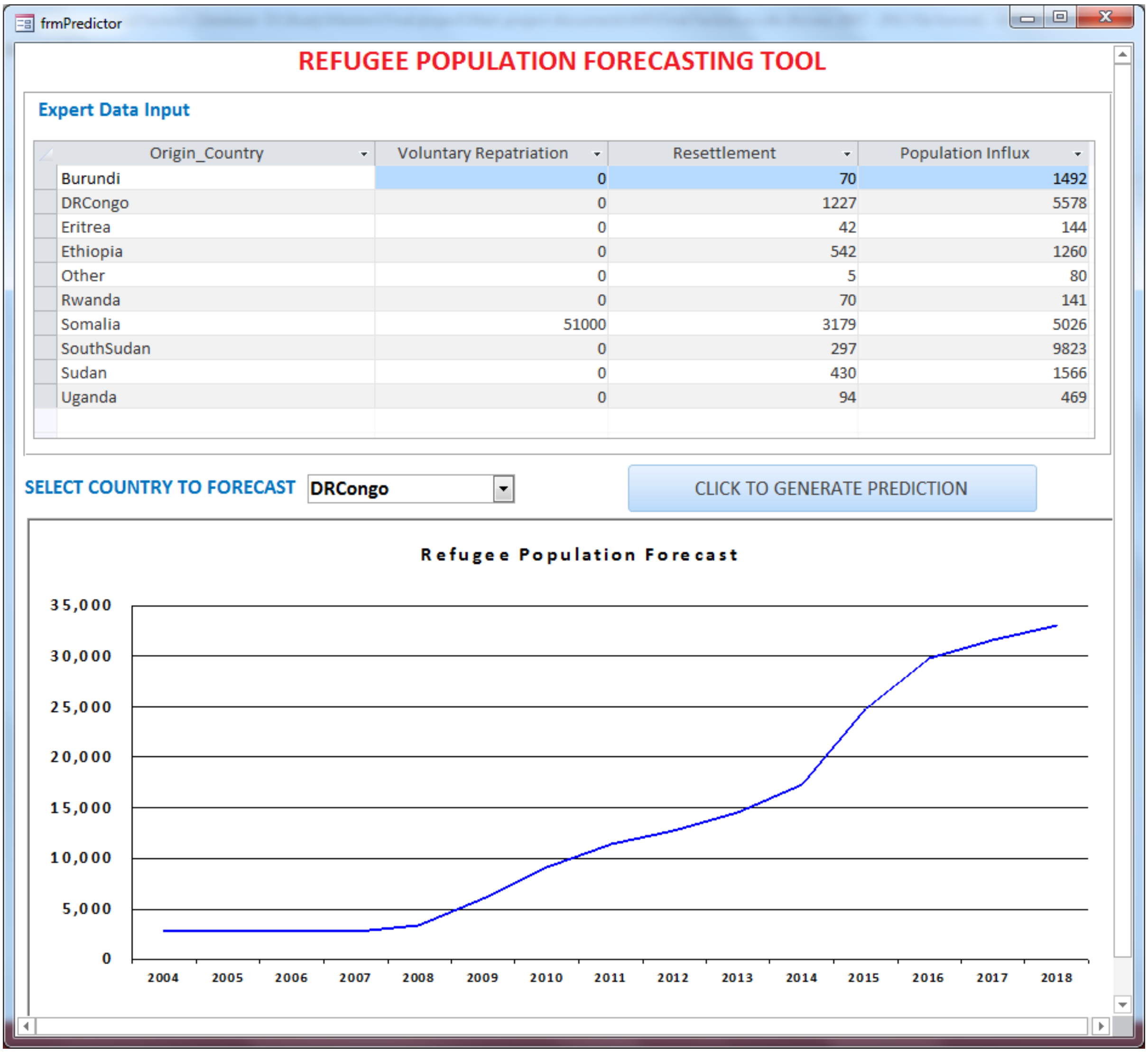
REFUGEE POPULATION FORECASTING TOOL

Expert Data Input

Origin_Country	Voluntary Repatriation	Resettlement	Population Influx
Burundi	0	70	1492
DRCongo	0	1227	5578
Eritrea	0	42	144
Ethiopia	0	542	1260
Other	0	5	80
Rwanda	0	70	141
Somalia	51000	3179	5026
SouthSudan	0	297	9823
Sudan	0	430	1566
Uganda	0	94	469

SELECT COUNTRY TO FORECAST: Burundi

Figure 24: Forecasting the Burundi refugee population



frmPredictor

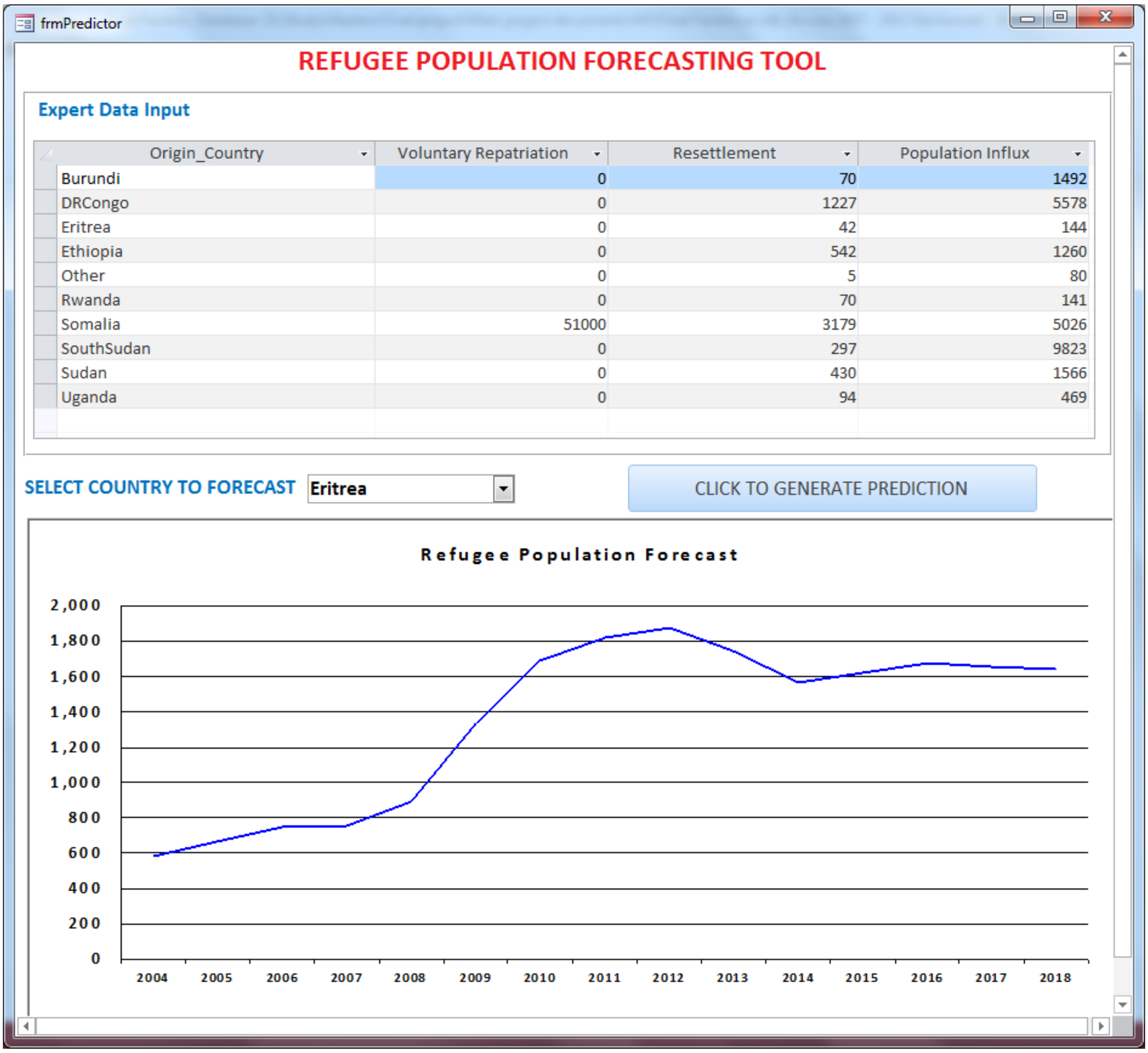
REFUGEE POPULATION FORECASTING TOOL

Expert Data Input

Origin_Country	Voluntary Repatriation	Resettlement	Population Influx
Burundi	0	70	1492
DRCongo	0	1227	5578
Eritrea	0	42	144
Ethiopia	0	542	1260
Other	0	5	80
Rwanda	0	70	141
Somalia	51000	3179	5026
SouthSudan	0	297	9823
Sudan	0	430	1566
Uganda	0	94	469

SELECT COUNTRY TO FORECAST: DRCongo

Figure 25: Forecasting the Congolese refugee population



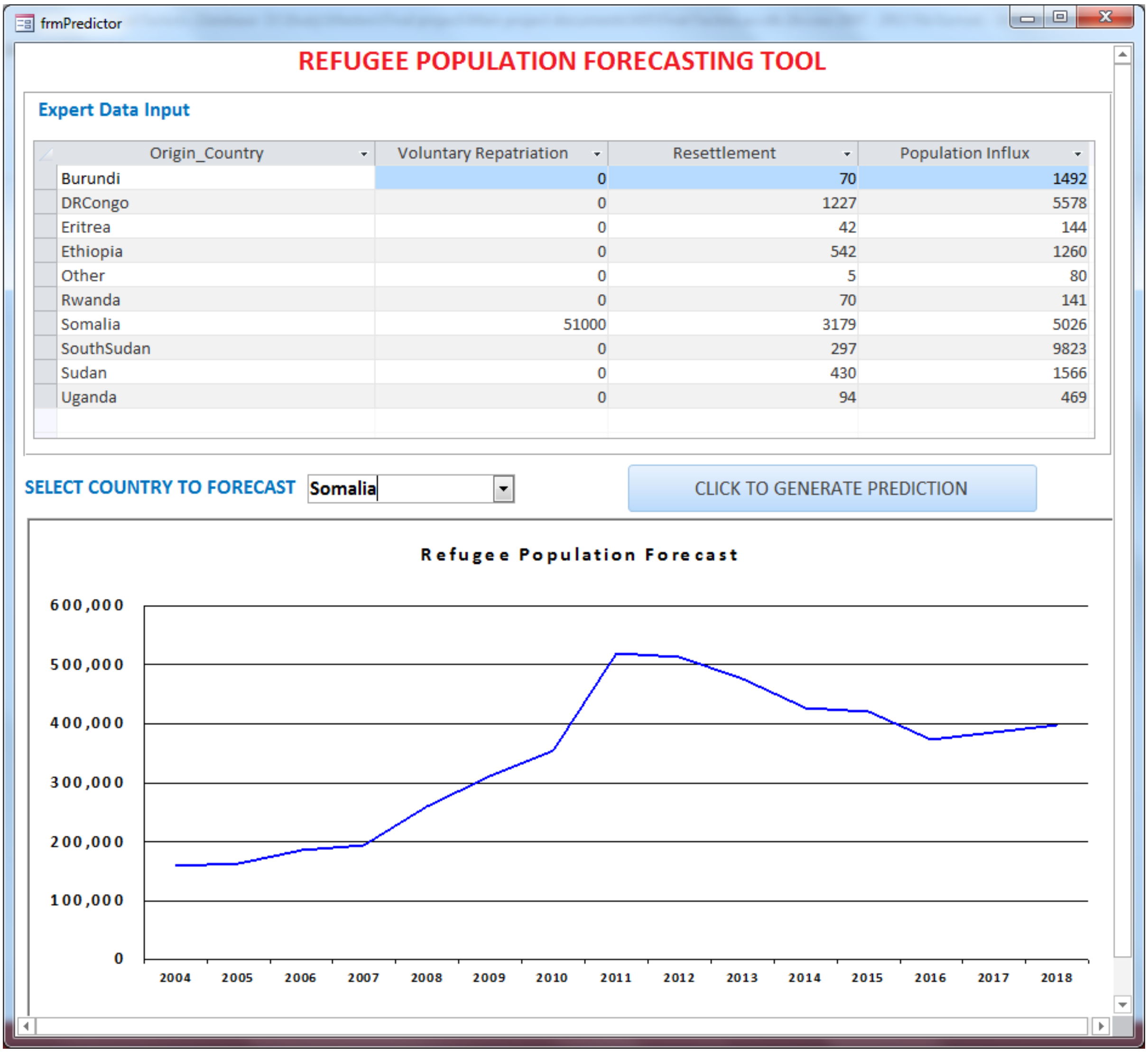
REFUGEE POPULATION FORECASTING TOOL

Expert Data Input

Origin_Country	Voluntary Repatriation	Resettlement	Population Influx
Burundi	0	70	1492
DRCongo	0	1227	5578
Eritrea	0	42	144
Ethiopia	0	542	1260
Other	0	5	80
Rwanda	0	70	141
Somalia	51000	3179	5026
SouthSudan	0	297	9823
Sudan	0	430	1566
Uganda	0	94	469

SELECT COUNTRY TO FORECAST: Eritrea

Figure 26: Forecasting the Eritrean refugee population



frmPredictor

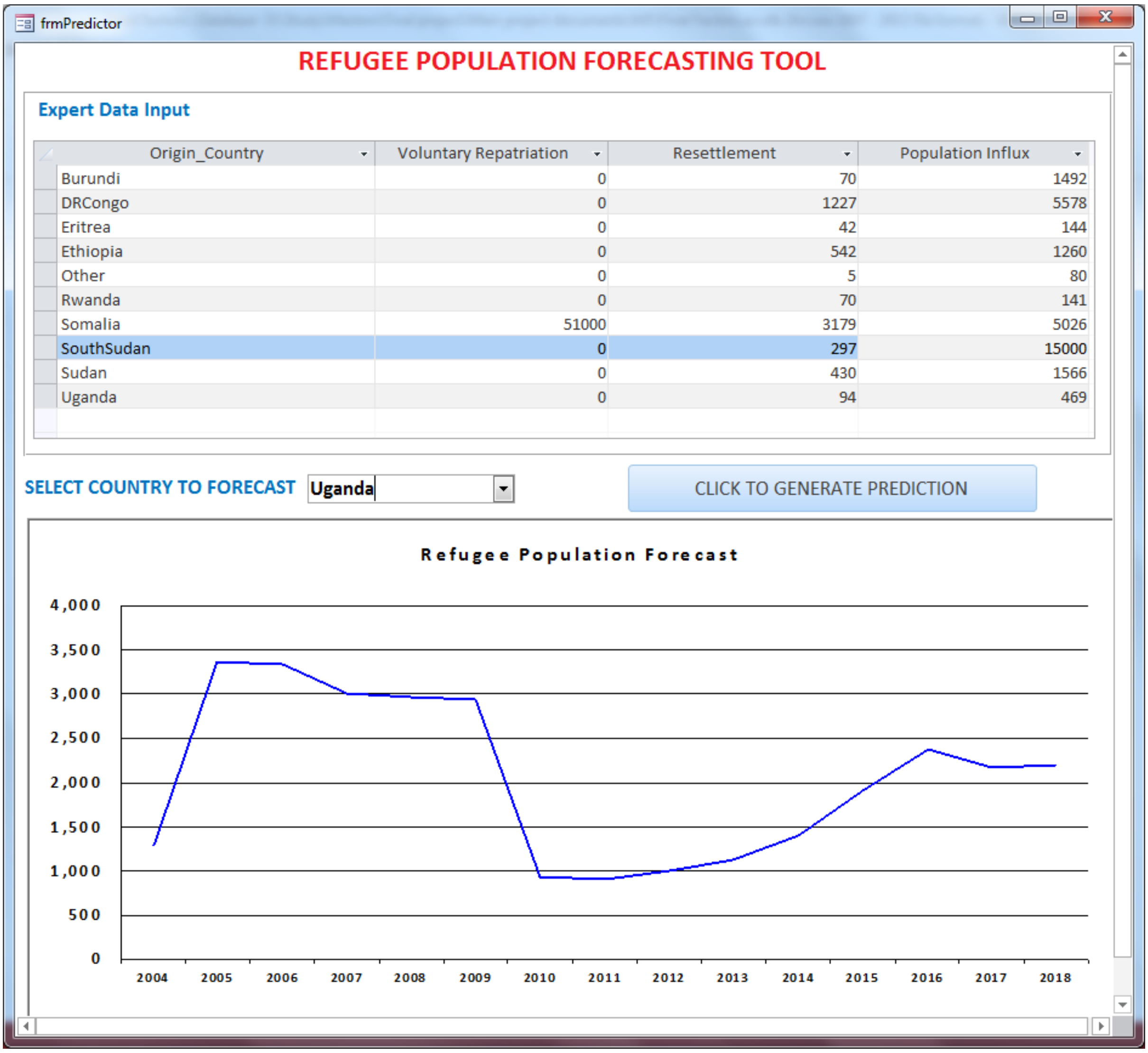
REFUGEE POPULATION FORECASTING TOOL

Expert Data Input

Origin_Country	Voluntary Repatriation	Resettlement	Population Influx
Burundi	0	70	1492
DRCongo	0	1227	5578
Eritrea	0	42	144
Ethiopia	0	542	1260
Other	0	5	80
Rwanda	0	70	141
Somalia	51000	3179	5026
SouthSudan	0	297	9823
Sudan	0	430	1566
Uganda	0	94	469

SELECT COUNTRY TO FORECAST: Somalia

Figure 27: Forecasting the Somali refugee population



REFUGEE POPULATION FORECASTING TOOL

Expert Data Input

Origin_Country	Voluntary Repatriation	Resettlement	Population Influx
Burundi	0	70	1492
DRCongo	0	1227	5578
Eritrea	0	42	144
Ethiopia	0	542	1260
Other	0	5	80
Rwanda	0	70	141
Somalia	51000	3179	5026
SouthSudan	0	297	15000
Sudan	0	430	1566
Uganda	0	94	469

SELECT COUNTRY TO FORECAST: Uganda

Figure 28: Forecasting the Ugandan refugee population

4.3 <u>Deployment</u>

The evaluation performed in Table 2 proves the possibility of deployment of this prototype for population projection. The prototype provides a hybrid of two approaches to prediction of population; the population predicted by trained models is complemented with expert input which adds known information about the present and future to the trend analyzed from historical data.

<u>Monitoring and maintenance</u>

The generation of models is a once a year activity once new statistics are uploaded by UNHCR. Input from experts on the other hand will continually feed the prototype based on any new information that could affect the population. It will be good practice to create a schedule such as quarterly or every half year to review the consistency of the prediction against the actual population reported monthly to monitor the accuracy of the model.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1. Achievements

The aim of the proposed research is to use data mining techniques to forecast refugee population in Kenya. The plan to achieve this was through training time series models using historical refugee population data obtained from a public refugee population statistics portal (http://popstats.unhcr.org). The data was extracted and prepared for the modelling technique selected; different models were generated for each nationality due to the establishment that the each country of origin has its own unique circumstances that lead to the refugee crises. Models generated were run against historical refugee data and used to make the forecasts.

This prototype will be tested by the UNHCR to make population forecasts for the years 2017 to 2019 and following any further adjustments deployed in all operations worldwide for the generation of planning figures for annual budgeting and donor engagement meetings.

5.2. Contribution of the study

The following outlines the contributions of this research.

i. The research area; introduction of trend analysis for forecasting refugee population in Kenya – the current process flow for generating population forecasts explained in Appendix B includes a set of seven factors applied to the last reported population to forecast the subsequent year's population. This procedure has to be repeated for each year to be forecasted. The use of models introduced by this research performs a trend analysis of the population allowing a user to forecast a number of years at once. From the evaluations performed as well as comparison of the proposed vis-à-vis the current forecasting procedure, it is clear that the prediction performed by the models are more straight forward and also quite accurate against actual statistics.

ii. Emphasis on uncertainty management in forecasting – this research exposed certain nationalities which could not be forecast correctly through machine learning methods only, this was because the historical data alone was not enough to predict how the population will look like in future. Countries such as South Sudan and Somalia which had drastic changes in either negatively or positively introduced the need for expert input to fill in the gaps of unforeseen events which could not be picked up through training models.

5.3. Limitations of the study

Majority of the information and resources required by this project were readily available, trend analysis has been used in majority of finance and economics application areas for a long time and various models and methodologies developed. This provided a lot of content for this research paper and provided options which could be tested to decide on an appropriate approach. The scope for this project was easy to set as the objectives were clear and the expected output had the current process to compare with. The current procedure used for forecasting refugee population provided figures which were used to assess the correctness of the developed prototype; this was used with the reported population statistics in September 2016 used as a baseline.

Although the research has reached its aims, project planning and implementation was not without limitations, the following were noted at different stages of the project:

i. Amount of refugee registration data available for generation of forecasting model – historical statistics on refugees was available from as far back as the year 1971. On close assessment of the data as well as literature review it was revealed that the credibility of the data could only be assured from around the year 2000 when proper structures and systems for registration of refugees and asylum seekers were setup. The quality and accuracy of the generated models increases with the amount of data available for training, validation and testing phases of model generation (Borovicka et al, 2012).

ii. Availability of data to fully validate the accuracy of the forecast output – the actual validation of predictions provided by the developed prototype can only be seen when the end of 2016 statistics are produced in December 2016 which is the first forecast year. This limitation was managed through the acquisition of the most recent available statistical data reported in September 2016, in addition the forecast generated by this prototype was compared against the predictions done by the current procedure. During the model creation process, the evaluation was done carefully to ensure that the lowest possible Mean Absolute Percentage Error (MAPE) is achieved to ensure that the models are as accurate as possible, additionally the input data was split well into a training set and test set data in a manner to avoid overfitting.

5.4. Recommendations for future work

This research work opens up time series forecasting to the field of humanitarian response which for a long time has not been included in majority of data mining related research as noted by Halevi (2012). This is however improving with the humanitarian field slowly acknowledging the importance of science and innovation. Following literature review and the experience from this research project, the following were identified as recommendations for future work related to this project.

i. Hybrid approach to handling uncertainty

   The uncertainty of forecasts of the size and age structure of population depends on the uncertainty of forecasts of fertility, mortality, and international migration (De Beer, 2000). The Foresight Manual by the Global Center for Public Service Excellence (2013) introduces the importance of Expert Panels in filling the gaps in forecast data. Keilman (2001) adds the importance of Expert judgement when expected values and corresponding prediction intervals are hard to obtain by formal methods. This is just one method out of the three main methods outlined by Tayman (2009) as listed in section 3.4. Though the expert input was very useful to involve the different players in fields affecting population, it would be necessary to include a hybrid approach which includes expert input as well as statistical probability.

ii. Expand research on machine learning techniques to other sectoral aspects of humanitarian response to enhance preparedness.

   This research project focused on the refugee context of humanitarian response and specifically on refugee population forecasting. For an all-round approach to preparedness for humanitarian response, all sectors need to be able to predict when they will be required to activate their response. There is need for research targeted to the various clusters of humanitarian response such Food Security, Education, Health, Shelter etc. based on findings from this project, this would require consistent collection and storage of such information so that it can be used to train prediction models.

REFERENCES

Agrawal, R. K. R. A. (2013). An Introductory Study on Time Series Modeling and Forecasting. arXiv Preprint arXiv:1302.6613, 1302.6613, 1–68.

Akkihal, A. R. (2006). Inventory Pre-positioning for Humanitarian Operations. Engineering Systems Division, 1–109.

Anderson, D., Sweeney, D., Williams, T., Camm, J., & Martin, R. (2011). An Introduction to Management Science: Quantitative Approaches to Decision Making, 1–896.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40–79. http://doi.org/10.1214/09-SS054

Belloni, R. (2011). The Birth of South Sudan and the Challenges of Statebuilding. *Ethnopolitics*, *10*(3-4), 411–429. http://doi.org/10.1080/17449057.2011.593364

Berry, M., & Linoff, G. S. (2009). Data Mining Techniques : Theory and Practice Course Notes.

Bezemer, D. J. (2009). “No One Saw This Coming”: Understanding Financial Crisis Through Accounting Models. MPRA Paper, 15892(15892), 1–51. Retrieved from http://www.heterodoxnews.com/htnf/htn85/No one saw this coming.pdf

Bontempi, G. (2013). Machine Learning Strategies for Time Series Prediction. Ulb.Ac.Be. Retrieved from http://www.ulb.ac.be/di/map/gbonte/ftp/time_ser.pdf

Booth, H., 2004. On the importance of being uncertain: Forecasting population futures for Australia. People and Place, 12(2), pp.1–12.

Borovicka, T. et al., 2012. Selecting Representative Data Sets. *Advances in Data Mining Knowledge Discovery and Applications*, pp.43–70.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2010). WEKA Manual for Version 3-7-2. Interface.

Brockwell, P. J., & Davis, R. a. (2002). Introduction to Time Series and Forecasting , Second Edition.

Brownlee, J. (2013). How to Evaluate Machine Learning Algorithms - Machine Learning Mastery. Machine Learning Mastery. Retrieved 7 August 2016, from http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. CRISP-DM Consortium, 76.

De Beer, J., 2000. Dealing with uncertainty in population forecasting. Statistics Netherlands. Department of population, pp.1–40.

De Gooijer, J. G., Hyndman, R. J., Gooijer, J. G. De, & Hyndman, R. J. (2006). 25 Years of Time Series Forecasting. International Journal of Forecasting, 22(January), 2006. http://doi.org/10.1016/j.ijforecast.2006.01.001

Edelstein, H. A. (1999). Introduction to Data Mining and Knowledge Discovery. Data Mining and Knowledge Discovery Handbook (Vol. 2). http://doi.org/10.1007/978-0-387-09823-4_1

Fallis, A. (2013). No Title No Title. Journal of Chemical Information and Modeling (Vol. 53). http://doi.org/10.1017/CBO9781107415324.004

Fadiya, S. O., Saydam, S., & Zira, V. V. (2014). Advancing data for humanitarian needs. Procedia Engineering, 78, 88–95. http://doi.org/10.1016/j.proeng.2014.07.043

Goldstein-Rodriguez, R. (2016). UNHCR seeks ProGres in refugee registration. UNHCR. Retrieved 7 August 2016, from http://www.unhcr.org/news/latest/2004/9/4135e9aa4/unhcr-seeks-progres-refugee-registration.html

Halevi, G. (2012). Special Issue on Data. Research Trends, (30), 1–40. Retrieved from http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). Principles of data mining. Drug safety : an international journal of medical toxicology and drug experience (Vol. 30). http://doi.org/10.2165/00002018-200730070-00010

Handbook for Emergencies. (2007) (3rd ed.). Geneva, Switzerland. Retrieved from http://www.unhcr.org/472af2972.html

Hyndman, R. (2016). Forecasting: principles and practice. Otexts.org. Retrieved 7 September 2016, from https://www.otexts.org/fpp

Hyndman, R. (2016). Why every statistician should know about cross-validation. Robjhyndman.com. Retrieved 7 August 2016, from http://robjhyndman.com/hyndsight/crossvalidation/

Ismail, Y. (2016). Fingerprints mark new direction in refugee registration. UNHCR. Retrieved 7 August 2016, from http://www.unhcr.org/news/latest/2006/11/456ede422/fingerprints-mark-new-direction-refugee-registration.html

Keilman, N., 2001. Uncertain population forecasts. Nature, 412(6846), pp.490–491.

Larose, D. T. (2006). Data Mining Methods and Models. Data Mining Methods and Models. http://doi.org/10.1002/0471756482

Limbu, M., Wanyagi, L., Ondiek, B., Munsch, B., & Kiilu, K. (2015). Kenya Inter-agency Rapid Assessment Mechanism (KIRA): A Bottom-up Humanitarian Innovation from Africa. Procedia Engineering, 107(M), 59–72. http://doi.org/10.1016/j.proeng.2015.06.059

Lutz, W. et al., 2000. New Developments in the Methodology of Expert- and Argument-Based Probabilistic Population Forecasting. *Director*, p.p.22.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2008). Introduction to TIme Series Analysis and Forecasting. http://doi.org/10.1017/CBO9781107415324.004

Nelson, E. L., & Gregg Greenough, P. (2016). Chapter 49 - Geographic Information Systems in Crises. Disaster Medicine (2nd ed.). Elsevier Inc. http://doi.org/http://dx.doi.org/10.1016/B978-0-323-28665-7.00049-2

Olson, D. L. D., & Delen, D. (2008). Advanced data mining techniques. http://doi.org/10.1007/978-3-540-76917-0

Parise, S., Iyer, B., & Vesset, D. (2012). Four strategies to capture and create value from big data. http://iveybusinessjournal.com/. Retrieved 1 April 2016, from http://iveybusinessjournal.com/publication/four-strategies-to-capture-and-create-value-from-big-data/

Popstats.unhcr.org,. (2016). UNHCR Population Statistics - Data - Overview. Retrieved 20 January 2016, from http://popstats.unhcr.org/en/overview

Rckkenya.org,. (2016). Refugees, Asylum seekers and Returnees. Retrieved 20 January 2016, from http://www.rckkenya.org/index.php/facts-and-faqs/refugees-asylum-seekers-returnees

Reefke, H. (2010). Simulation of container traffic flows at a metropolitan seaport. Lecture Notes in Business Information Processing (Vol. 46 LNBI). http://doi.org/10.1007/978-3-642-12494-5_37

Rohanizadeha, S. and Moghadam, M. (2009). A Proposed Data Mining Methodology and its Application to Industrial Procedures. Journal of Industrial Engineering, pp.37-50.

Rohrer, B. (2016). How to choose algorithms for Microsoft Azure Machine Learning. Azure.microsoft.com. Retrieved 7 September 2016, from https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/

Simon, E. (2002). Forecasting foreign exchange rates with neural networks.

Smith, M. (2014). A Comparison of Time Series Model Forecasting Methods on Patent Groups.

Tamboli, A. (2010). An Introduction to Data Mining | BPM, Lean Six Sigma & Continuous Process Improvement | Process Excellence Network. [online] Processexcellencenetwork.com. Available at: http://www.processexcellencenetwork.com/lean-six-sigma-business-transformation/articles/an-introduction-to-data-mining [Accessed 19 Mar. 2016].

Tayman, J., 2009. Uncertainty in Forecasting.

Thearling, K. (2012). An Introduction to Data Mining. Journal of Postgraduate Medicine, 39(2), 105. http://doi.org/10.1016/j.cll.2007.10.008

Tobergte, D. R., & Curtis, S. (2013). No Title No Title. Journal of Chemical Information and Modeling (Vol. 53). http://doi.org/10.1017/CBO9781107415324.004

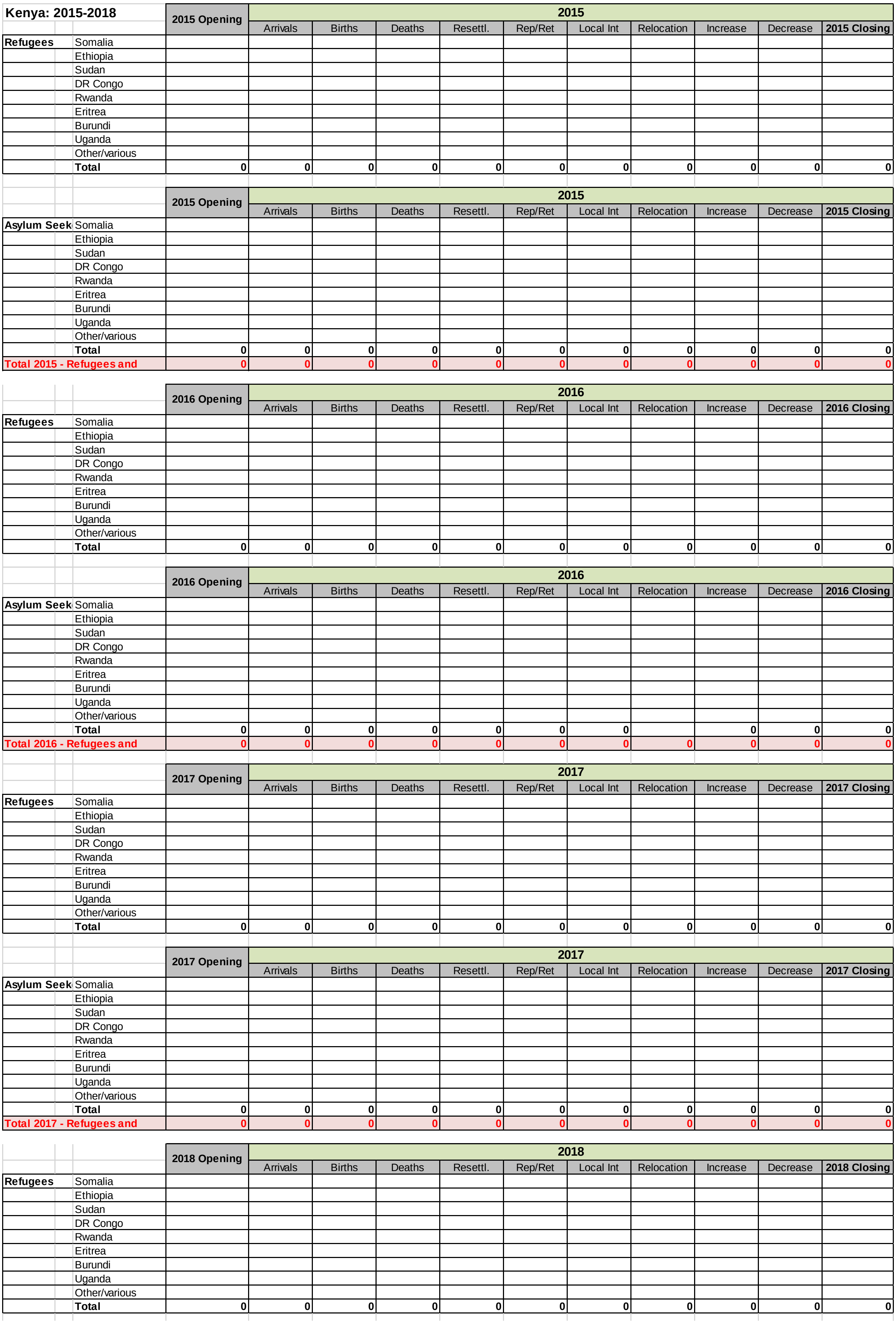
Unocha.org,. (2016). Big data and humanitarianism: 5 things you need to know | OCHA. Retrieved 26 February 2016, from http://www.unocha.org/top-stories/all-stories/five-things-big-data-and-humanitarianism

Whipkey, K., & Verity, A. (2015). Guidance for Incorporating Big Data into Humanitarian Operations, 42. Retrieved from http://digitalhumanitarians.com

Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. Annals of Physics (Vol. 54). http://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Zaki, M. J., & Meira, M. J. (2013). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

APPENDIX A: Refugee population planning figures forecasting template



Kenya: 2015-2018		2015 Opening	2015									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2015 Closing
Refugees	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

		2015 Opening	2015									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2015 Closing
Asylum Seek	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Total 2015 - Refugees and		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

		2016 Opening	2016									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2016 Closing
Refugees	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

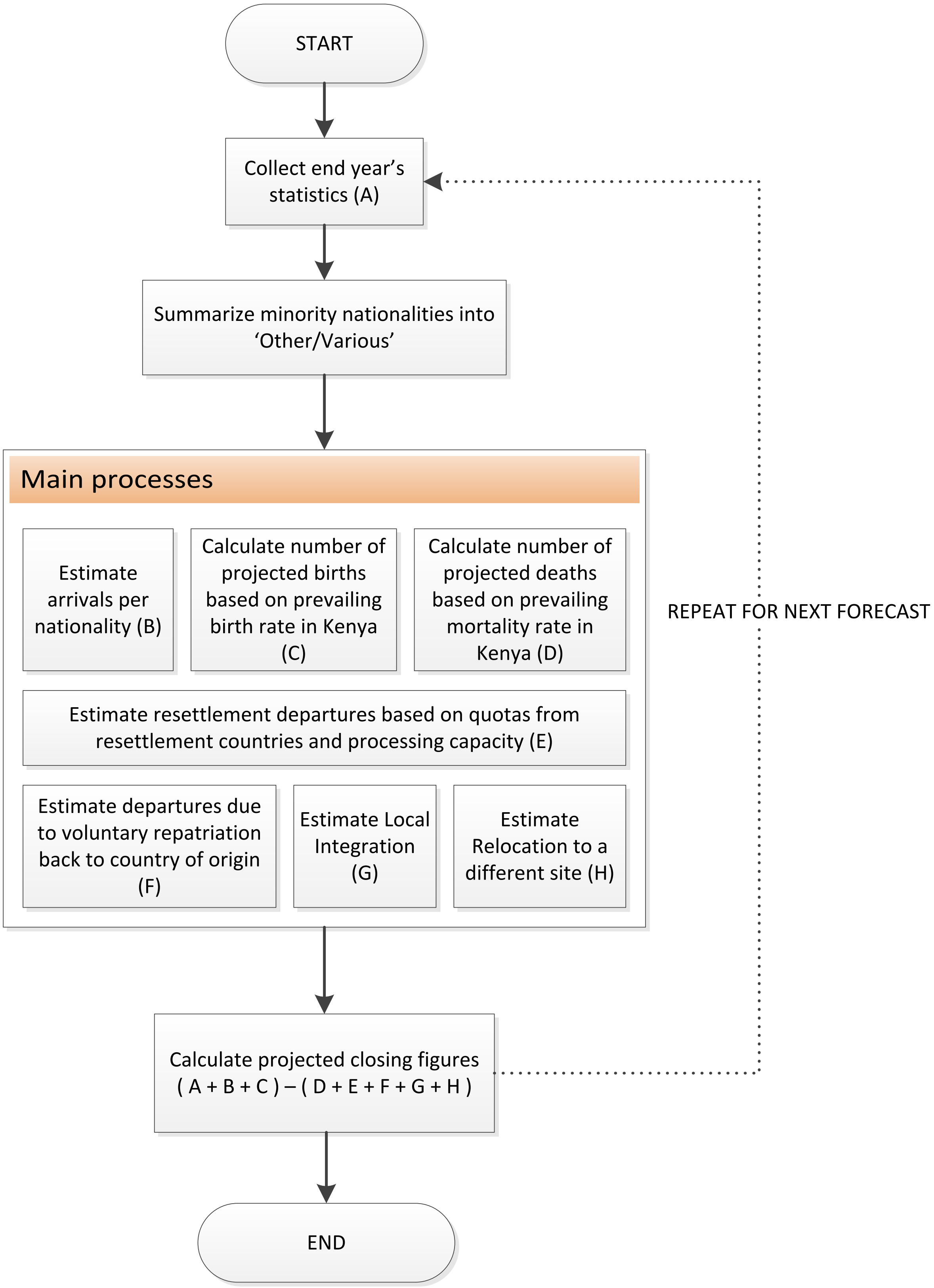
		2016 Opening	2016									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2016 Closing
Asylum Seek	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**		**0**	**0**	**0**
Total 2016 - Refugees and		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

		2017 Opening	2017									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2017 Closing
Refugees	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

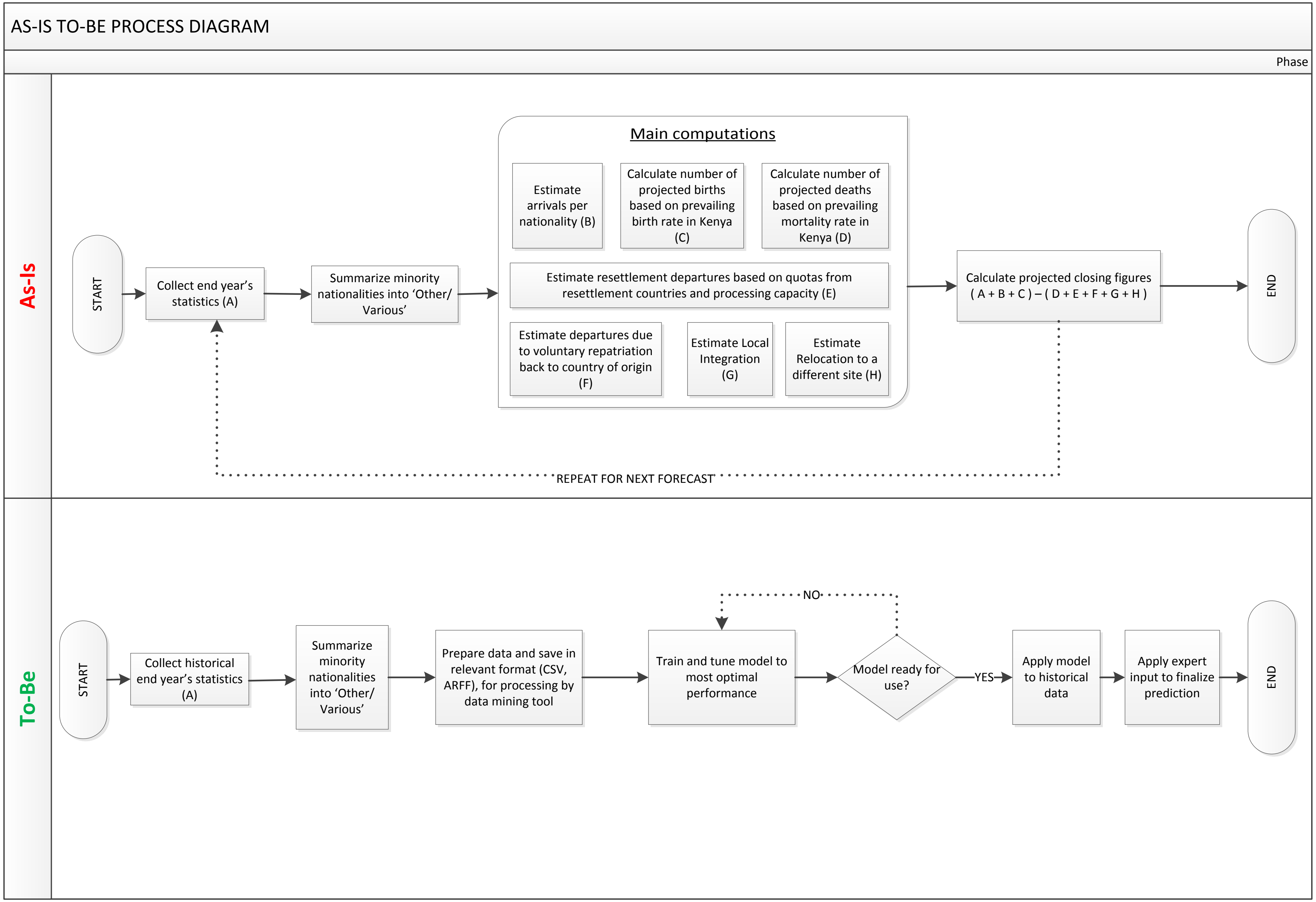
		2017 Opening	2017									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2017 Closing
Asylum Seek	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Total 2017 - Refugees and		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

		2018 Opening	2018									
			Arrivals	Births	Deaths	Resettl.	Rep/Ret	Local Int	Relocation	Increase	Decrease	2018 Closing
Refugees	Somalia											
	Ethiopia											
	Sudan											
	DR Congo											
	Rwanda											
	Eritrea											
	Burundi											
	Uganda											
	Other/various											
	Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

APPENDIX B: Flow chart – Current procedure for forecasting refugee population


START
Collect end year's statistics (A)
Summarize minority nationalities into 'Other/Various'
Main processes
Estimate arrivals per nationality (B)
Calculate number of projected births based on prevailing birth rate in Kenya (C)
Calculate number of projected deaths based on prevailing mortality rate in Kenya (D)
Estimate resettlement departures based on quotas from resettlement countries and processing capacity (E)
Estimate departures due to voluntary repatriation back to country of origin (F)
Estimate Local Integration (G)
Estimate Relocation to a different site (H)
Calculate projected closing figures (A + B + C) – (D + E + F + G + H)
REPEAT FOR NEXT FORECAST
END

APPENDIX C: Process modelling – As-Is To-Be diagram


AS-IS TO-BE PROCESS DIAGRAM
Phase
As-Is
START
Collect end year's statistics (A)
Summarize minority nationalities into 'Other/ Various'
Main computations
Estimate arrivals per nationality (B)
Calculate number of projected births based on prevailing birth rate in Kenya (C)
Calculate number of projected deaths based on prevailing mortality rate in Kenya (D)
Estimate resettlement departures based on quotas from resettlement countries and processing capacity (E)
Estimate departures due to voluntary repatriation back to country of origin (F)
Estimate Local Integration (G)
Estimate Relocation to a different site (H)
Calculate projected closing figures (A + B + C) – (D + E + F + G + H)
END
REPEAT FOR NEXT FORECAST
To-Be
START
Collect historical end year's statistics (A)
Summarize minority nationalities into 'Other/ Various'
Prepare data and save in relevant format (CSV, ARFF), for processing by data mining tool
Train and tune model to most optimal performance
NO
Model ready for use?
YES
Apply model to historical data
Apply expert input to finalize prediction
END

SAMPLE CODE

```
Option Compare Database


Private Sub cmdGenerate_Click()

[Forms]![frmPredictor]![cmbCountry].SetFocus

graphPrediction.RowSource = "TRANSFORM Sum(qryGeneratedPrediction.ForecastIndividuals) AS
SumOfForecastIndividuals SELECT qryGeneratedPrediction.Yr FROM qryGeneratedPrediction WHERE
(((qryGeneratedPrediction.Origin_Country)='" & [Forms]![frmPredictor]![cmbCountry] & "')) GROUP BY
qryGeneratedPrediction.Yr PIVOT qryGeneratedPrediction.Origin_Country;"

Me.graphPrediction.Requery


End Sub
```

```
SELECT Data.ID, Data.Source_ID, Source.DataSource, Year([Reporting_Year]) AS Yr,
Data.CountryOfOrigin_ID, CountryOfOrigin.Origin_Country, Data.Num_Individuals

FROM (Data LEFT JOIN CountryOfOrigin ON Data.[CountryOfOrigin_ID] = CountryOfOrigin.[ID]) LEFT JOIN
Source ON Data.[Source_ID] = Source.[ID];
```

```
SELECT ForecastData.DataSource, ForecastData.Yr, ForecastData.Origin_Country,
ForecastData.Num_Individuals, IIf([DataSource]="Forecast",[Num_Individuals]-[tmpExpertData]![VolRep]-
[tmpExpertData]![Resettlement]+[tmpExpertData]![PopulationInflux],[Num_Individuals]) AS
ForecastIndividuals

FROM ForecastData INNER JOIN tmpExpertData ON ForecastData.Origin_Country =
tmpExpertData.OriginCountry

ORDER BY ForecastData.Yr;
```