Human Displacement in South Sudan

Primary Topic: Data Mining, Secondary Topic: Data Visualisation
Course: 2018-1A – Group: 15 – Submission Date: 2018-02-06

Hau-Ben Benjamin (Ben) Shih
University of Twente
h.b.shih@student.utwente.nl

Yi-Hsiou Hsu
University of Twente
y.hsu-1@student.utwente.nl

ABSTRACT

In several parts of the world, migration and forced displacement due to local conflicts, social persecution, gender orientation, hunger and climate change have been detected in increasing numbers. There are an estimated of 25.4 million displaced refugees worldwide now, and South Sudan, the focus of this paper, accounts for 10% of those refugees [1].

Predicting refugee movement is important for organizations like the United Nations Refugee Agency (UNCHR); an accurate prediction can help save refugee lives by allowing relief organisations to conduct better-informed allocation of humanitarian resources [2]. In this project, we will use the data provided by UNHCR and apply data mining techniques to look for ways to help UNHCR estimate or predict the total growing number of refugees and find the main factors that causes refugees' movements.

The main algorithms used was K-means algorithm; with the support of Decision tree classifier and Apriori algorithm to help our final interpretation of data. The report concludes by discussing what we derived from the provided data, the predictable and non-predictable factors and recommendations for future works.

KEYWORDS

Human displacements, South Sudan, UNHCR, Machine Learning, Data Mining, Data Visualisation.

1 INTRODUCTION

After the Rwandan Genocide of 1994, when the world stood by and watched the slaughter of 800,000 people, the United Nations stated that the world has an obligation to intervene and prevent ethnic cleansing. The UN said it would never allow that to happen again – but it has. In November 2016, the U.N. commission of human rights visited South Sudan, the world's youngest country. They found a conflict marked by mass slaughter and what they described as a "warped environment" where the rape of women and girls has become normal [3]. Since the civil war of South Sudan broke out in December 2013, more than 400,000 people have been killed [4]. The situation in South Sudan reached an unprecedented level in 2017 as more than 2.1 million people have been forced to flee their homes [1]. For people living in the country, life has became a total nightmare. The country gained its independence just eight years ago in a move that was supposed to bring peace to an area that had known only war. With the total numbers of displaced refugees rising, forced displacement is the most urgent issue that the United Nations Refugee Agency (UNCHR) must deal with as the current displacement situation reaches unprecedented levels and creates a grave humanitarian crisis [5].

As specified for this project, we will mainly use the provided datasets to discuss two objectives:

- How can machine learning assist in predicting human displacement in the country of South Sudan?
- What are the most influential factors that affect the displacement of persons of concern (POC)?

2 METHODOLOGY

For the entire project, we will follow cross-industry standard process for data mining (CRISP-DM) to learn from the provided database. A detail explanation of CRISP-DM can be find in Appendix A. In the following section, we will explain what we did in each step of the process in detail.

2.1 Background Understanding

Renewed Efforts Against Child Hunger and undernutrition (REACH), a joint initiative with the United Nations Operational Satellite Applications Program, published a report in 2017 with key information about the main push-and-pull factors of human displacements within South Sudan. For example, the report found that most refugees suggested that they will only return to South Sudan after definitive peace is reached in the country; despite this, many people have returned to the country due to the difficult conditions in refugee camps. The main factor pushing people to leave was the civil war, without any doubt. However, more studies are required to find out which countries refugees migrated to and why [6].

2.2 Data Understanding

We began our project with 21 datasets provided by UNHCR. The datasets were separated into five main categories, which are described below.

1. ACLED_south-sudan (Violent and Non-Violent events): The ACLED datasets tracks a range of violent and non-violent events that have happened in South Sudan since October 2011. The data contain specific information on the dates, locations, group names, interaction types, event types and reported fatalities as well as contextual notes; the details of different event types are explained in Appendix B.

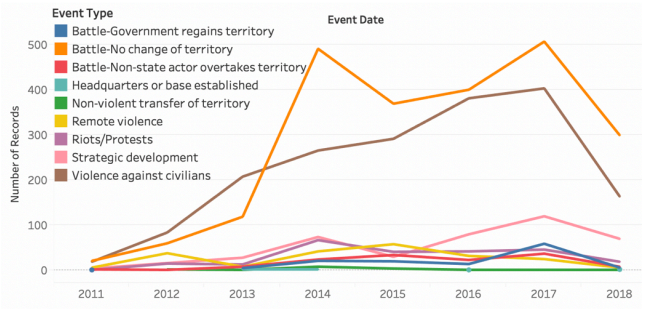

Figure 1 The trend of number of different kind of events in each year.

From figure 1 above we learn that the number of violent events has been increasing since 2012 and reached a maximum number of violent and non-violent events between 2013–2014 and 2016–2017. Most of these events were 'battles to take over territory' and 'violence against civilians'. Moreover, the number of strategic developments have been increasing recently. As defined by ACLED, strategic development is defined as non-violent events that have the potential to trigger future events; these events include the arrests of key political figures, rallies, peace talks, etc.

2. Refugees from South Sudan: This dataset records the total number of refugees that have been displaced in foreign countries every month since 2014. As seen in figure 2, the total number of refugees from South Sudan has been increasing since the beginning of the rebellion in 2014. The number steadily rose until a compromise peace agreement fell apart in July 2016. The fight spreads to the previously safe haven, Upper Nile in October 2017, which caused 6 million people to face starvation. The new peace agreement was signed in October 2018, which could likely be the reason for the significant drop in numbers in 2018.

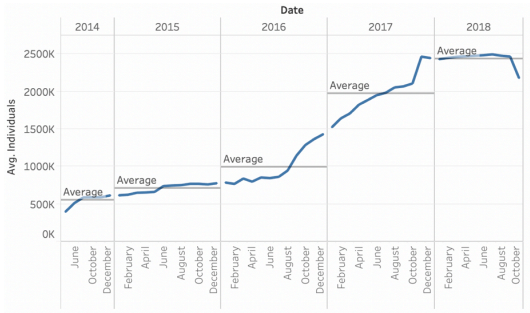

Figure 2 The total number of displaced South Sudanese in each year

3. Refugees by Country: Several datasets have been grouped in this category. In these datasets, we can check the total number of South Sudanese that have migrated to Kenya, Democratic Republic of Congo (DRC), Ethiopia, Sudan and Uganda. The number has been updated every month since 2014. The visualisation of these datasets can be seen in Appendix C, most refugees currently live in Sudan, Uganda and Ethiopia, and the total number of refugees in Uganda and Sudan has increased since 2016 and 2017, respectively.

4. Refugees by Country in Detail: These datasets represent a more detailed version of the category mentioned above. Instead of recording the total number of South Sudanese in a country, it records the province, gender, age, etc. where the refugees migrate to. However, these datasets are more difficult to study as they are updated inconsistently.

5. wfp_food_prices_south-sudan: The dataset here records the price of daily essential items for each month. The main categories that are recorded are cereals and tubers, pulses and nuts, non-food (i.e. fuel), oils and fats, miscellaneous food, and meat, fish and eggs. It also contains the location of where the price was recorded.

As seen from Appendix D, the price of non-food items has increased significantly since 2017. We assume that this is mainly because South Sudan is one of the largest producers of oil and the civil war has affected the production of their oil fields.

2.3 Data Preparation

A few relations were identified among the datasets. For example, the increase of violent and non-violent events can be related the increase of South Sudanese displacements, and the region in which the events happened could be related to the number of South Sudanese that migrate to the nearest country in the region. In this section, we focus on merging the datasets so that we can generate an overview model for prediction. To avoid data redundancy, we eliminated some data fields that were not useful for our intended analysis. A set of processes was performed on the datasets before we used them to train the prediction models:

ACLED Violent and Non-Violent datasets: We kept the date, event types and region (province) in which the event happened and omitted the rest. There are still some important columns, such as actor type and interaction type, but we found most of them can be derived from event type, so we decided to eliminate them. After columns were selected, we transformed the data into a numerical data format. We grouped the table by year and month and converted the columns, such as event types and region, into separate columns. The final format of this table indicates the number of different violent and non-violent event types (with type names) that happened each month, plus the number of events that happened in each province.

Refugees from South Sudan / Refugees by Country: We found the structure for these two categories of datasets are similar so we merged them together in this stage. Before merging them, we also discovered that we were not interested in the total number of refugees. Instead, we subtracted the total number in each month

from the previous month to find out the number of refugees that moved in or out of each country in that month. An issue that we discovered after merging them is that there are some missing values in the datasets. To avoid problems, we used the padding technique[1] to handle the missing values.

wfp_food_prices_south-sudan: For this dataset, we kept the categories and the prices only; the rest of the fields were eliminated. We followed the same process again to convert the dataset structure to the same format as the others.

Since all the datasets were formatted into a list of records per month, we could join all the datasets together by grouping them their year and month. The final structure of this data enable us to compare the correlations between different values among the datasets. The final structure of the table is shown in Appendix E.

2.4 Modeling

Following the data structure that we described in Appendix E, we found that K-Means clustering could be useful to break the data into different groups that help us identify what happened to each group of clusters. To provide more confidence in the predicted model, we also performed several algorithms, such as Apriori, Decision tree, to support our predictions. A time-series model was also trained to forecast the future number of displaced South Sudanese. However, we believed that the model wasn't accurate enough as more considerations are needed to prepare the time-series model. Due to this, we have placed our time-series model in Appendix F for reference only.

2.4.1 K-Means Clustering on Full Dataset

K-means clustering is an unsupervised machine learning algorithm that can group datasets into user-specified numbers (k) of groups without prior knowledge. The algorithm for K-means clustering can be found in Appendix G. We found K-means algorithms useful for our dataset since we did not have any prior knowledge of our dataset and we wanted to check how fields in the datasets related to each other and how the full dataset could be grouped in different ways.

As seen in the figure 3, the algorithm divided our dataset into three clusters, which we noted as follows:

Cluster 0: From looking at the number of refugees that moved out of the country, we defined this group as the group of people who moved during the "Peaceful Period". People were coming back to South Sudan from other countries during this period (The negative number in refugees_move_out). The prices of everyday essentials were at their maximum and fuel prices (non-food) were growing aggressively. There were fewer violent and non-violent events in most parts of the country, and battles and violence against civilians were less. However, the number of strategic developments was also high during this period which could cause greater conflicts in the future.

[1] Padding: Propagate the last valid observation forward to next missing value

Cluster 1: This cluster represents be the hardest period for refugees, this is the 'War Starting Period'. People were moving out of the country significantly, as violent and non-violent events were generally occurring in the politically central areas (capital cities with government authority). The number of violent events against civilians and battles around the country increased and the price of different items increased. Most people migrated to Uganda and Sudan. The prices were a bit above-average at this point.

Cluster 2: We defined this group as the "During War Period". People were constantly moving out to neighboring countries because of the consistent violent and non-violent events that happened across the country. During this period, the price of most items was at their lowest.

Final cluster centroids:

Attribute	Full Data (49.0)	Cluster# 0 (8.0)	1 (18.0)	2 (23.0)
Central.Equatoria	17.8367	17.75	26.2778	11.2609
Eastern.Equatoria	5.1224	3.125	7.7778	3.7391
Gogrial	0	0	0	0
Jonglei	8.8163	5.625	12.6111	6.9565
Lakes	4.9796	4.375	5.8333	4.5217
Northern.Bahr.el.Ghazal	1.4286	0.5	1.4444	1.7391
Unity	10.4286	6.5	11.8889	10.6522
Upper.Nile	11.2653	3.5	14.3889	11.5217
Warrap	2.7755	2.5	3.1667	2.5652
Western.Bahr.el.Ghazal	7.0816	8.5	6.7778	6.8261
Western.Equatoria	6.5306	3.25	8	6.5217
Battle.Government.regains.territory	2.0816	0.75	3.5556	1.3913
Battle.No.change.of.territory	33.5306	28.125	41.7778	28.9565
Battle.Non.state.actor.overtakes.territory	2.1429	1	3	1.8696
Headquarters.or.base.established	0.0612	0	0.1111	0.0435
Non.violent.transfer.of.territory	0.1429	0.125	0.0556	0.2174
Remote.violence	2.7347	0.5	1.7778	4.2609
Riots.Protests	3.3673	1.75	3.5	3.8261
Strategic.development	6.1429	7.5	9.6111	2.9565
Violence.against.civilians	26.0612	15.875	34.7778	22.7826
refugees_move_out	26684.9388	-34910.375	68284.6667	15553.087
out_to_DRC	1888.2449	689.25	3320.2222	1184.6087
out_to_Ethiopia	4532.6531	-5865.125	8304.3333	5197.5217
out_to_Kenya	1137.6327	89.375	1340.3889	1343.5652
out_to_Sudan	14776.9796	-145.625	20020.6667	15863.6957
out_to_uganda	16399.3061	-28450.5	43132.6111	11077.5217
cereals.and.tubers	159.2342	339.198	231.1777	40.3345
non.food	542.1723	2844.9497	181.7969	23.2391
oil.and.fats	173.5695	420.778	236.8772	38.0387
pulses.and.nuts	182.4508	471.049	240.4775	36.6566

Figure 3 K-means clustering on the dataset we merged

Based on these results, it seems like the number of violence against civilians and strategic development affect the migration situation the most, but it is difficult to define how the number of violent / non-violent events in each region correlate to the country to which refugees migrate.

2.4.2 K-Means Clustering on ACLED Event Types and Number of Displaced Refugees.

In this part, we will run k-means to support our previous finding on the correlation between event types and the number of move outs from the country. With this portion of the attributes, four clusters were identified. The results are shown in figure 4.

From the clusters we identified, we can see that the groups are likely to be separated by the number of strategic development and violence against civilians, with cluster 0 having the highest number of both event types and cluster 1 having the lowest number of violence against civilians but the second highest number of strategic development. From section 2.2, we learned that the number of displaced refugees often decreases when a

peace agreement is signed and increases when the agreement fell, then steadily increase during the war. With this background, we could interpret the data as cluster 1 is likely to appear before cluster 0, and cluster 0 will lead to situations in cluster 2 and 3. In this way, we could find that the number of violence against civilians is one big factor that trigger refugees to move. Also, the number of strategic development might be the trigger point of the next war.

Final cluster centroids:

Attribute	Full Data (49.0)	Cluster# 0 (10.0)	1 (18.0)	2 (14.0)	3 (7.0)
Battle.Government.regains.territory	2.0816	2.7	0.7222	1.5	5.8571
Battle.No.change.of.territory	33.5306	44.6	31.4444	27.4286	35.2857
Battle.Non.state.actor.overtakes.territory	2.1429	4.1	1.1111	0.7857	4.7143
Headquarters.or.base.established	0.0612	0	0.1111	0.0714	0
Non.violent.transfer.of.territory	0.1429	0	0	0	1
Remote.violence	2.7347	1.6	0.8889	5.7857	3
Riots.Protests	3.3673	3.6	3.2222	3.1429	3.8571
Strategic.development	6.1429	10.6	6.5	2.8571	5.4286
Violence.against.civilians	26.0612	44.4	19.6111	20.9286	26.7143
refugees_move_out	26684.9388	97257.3	-9501.5	22493.4286	27301.1429

Figure 4 K-means clustering on violence type and human displacement numbers

2.4.3 Categorized Data

To provide greater confidence in the results, in the following sections we normalize the dataset so that each dimension lies between 0 and 1, so they have equal weight in the clustering process. To make the presentation of the normalized table cleaner, we have transformed each value into a nominal value. All values in the table are being transformed into either "Low", "Average" or "High". If the value is lower than 33% of the entire value of the column, it is categorized as "Low", and if the value is higher than 66% of the entire value in the column, it is categorized as "High". If the value is between High and Low, it is categorized as "Average".

K-Means on the Categorized Full Dataset

We ran K-means algorithms on the entire dataset again to check if we could identify any changes after we transformed the data. A few variables were ignored as there was not a way to separate them into three categories as their variances were too small.

By taking a suggestion from the canopy algorithm[2], three clusters were being identified. The result is shown in figure 5. From this model, we learn to identify a pattern of refugee movement. From the high volume of people leaving South Sudan in cluster 0, we can see that most of them did not choose to migrate to Sudan (which hosts the second-largest number of South Sudanese refugees). The volume of violent and non-violent events in the nearest region to Sudan, Upper Nile, was not very high during this period. However, in cluster 2, the period during war, we see that the number of people going into Sudan grew, just as the number of violent and non-violent events in Upper Nile grew. In summary, we were able to recognise that as the war broke out in the central political areas (i.e. the capital city) in the south part of the country, most people moved to Uganda since it is the nearest country. As the war continued, the battles moved towards the north; the reason for this could be that factions were battling for the oilfields in the north-central part of the country. When the battle started to move towards the north, more violent and non-violent events occurred; as a results, people in those areas preferred to move to the nearest country for safety.

Final cluster centroids:

Attribute	Full Data (49.0)	Cluster# 0 (25.0)	1 (9.0)	2 (15.0)
Central.Equatoria	High	High	Average	Low
Eastern.Equatoria	High	High	Low	Low
Jonglei	Average	High	Average	Average
Lakes	High	High	Low	High
Unity	Average	Average	Average	High
Upper.Nile	High	Low	Average	High
Warrap	High	High	High	Average
Western.Bahr.el.Ghazal	Average	Average	High	Low
Western.Equatoria	High	High	Average	High
Battle.No.change.of.territory	Average	High	Average	Low
Battle.Non.state.actor.overtakes.territory	High	High	Low	Average
Remote.violence	High	Low	High	High
Riots.Protests	High	Average	High	High
Strategic.development	Average	High	Average	Low
Violence.against.civilians	Average	High	Average	Average
refugees_move_out	High	High	Low	Average
out_to_DRC	Average	Average	Average	Average
out_to_Ethiopia	High	High	Low	Average
out_to_Kenya	High	Low	Average	High
out_to_Sudan	High	Low	Average	High
out_to_uganda	Average	High	Average	Average
cereals.and.tubers	High	High	Average	Low
non.food	High	High	Average	Low
oil.and.fats	High	High	Average	Low
pulses.and.nuts	High	High	Average	Low

Figure 5 K-means clustering on categorized data

2.4.4 Decision Tree to Predict Number of Displaced Refugees

By using Decision tree, although the accuracy of the prediction model was barely above 50%, we identified the growth of violent/non-violent events in some particular region could be straightly related to the number of people moving out in that period. From the root, we can still find out violence against civilians is still the main factor why people move away from their home. An addition to this, the total number of violent/non-violent events in a few provinces from the west side, Western Bahr el Hazal and Western Equatoria, and Lakes could be a critical factor of predicting how many people will move out from South Sudan. From the tree, we were not able to distinguish the key points that cause this results, but we definitely recommend to work on it in the future to find out the detail in these provinces.

```
Violence.against.civilians = Average
|   Northern.Bahr.el.Ghazal = Average
|   |   Battle.Government.regains.territory = Average
|   |   |   Strategic.development = Average: Low (3.0/1.0)
|   |   |   Strategic.development = High: Low (2.0)
|   |   |   Strategic.development = Low: Average (5.0)
|   |   Battle.Government.regains.territory = High: Low (3.0/1.0)
|   Northern.Bahr.el.Ghazal = High: High (5.0/1.0)
Violence.against.civilians = High
|   Eastern.Equatoria = Average: High (1.0)
|   Eastern.Equatoria = High: High (12.0/1.0)
|   Eastern.Equatoria = Low: Average (4.0/1.0)
Violence.against.civilians = Low
|   Western.Bahr.el.Ghazal = Average: Average (5.0/1.0)
|   Western.Bahr.el.Ghazal = High: Low (2.0)
|   Western.Bahr.el.Ghazal = Low: Low (7.0/1.0)
```

Figure 6 Results from the decision tree for predicting refugee's movements

2.4.5 Decision Tree to Predict Fuel Price

As stated before, South Sudan owns one of the largest oil fields in the world. As a result, we believe the number of violent/non-violent events are related to the price of the fuel. The result of

[2] The canopy algorithms is an unsupervised pre-clustering algorithms that is often used to calculate the number of cluster that the k-means should generate. [7]

running Decision tree algorithm to predict this topic is shown in Appendix H. The accuracy of this tree was about 70%. We can find out some interesting provinces that are quite correlated with the price of the fuel. Central Equatoria, Unity and Lakes for example, if the number of violent/non-violent events increases, the fuel price will be increased as well. When the number of events in Central Equatoria, Unity and Lakes, it probably means that the fuel prices are low as well. This could be used as a signal for predicting future fuel prices.

2.4.6 Apriori Algorithm

The final machine learning algorithm we used to predict our data is the Apriori algorithm. The Apriori is often used to find associations between variables. We tried Apriori algorithms to find out any associations between our classes. The results from the Apriori algorithms is shown in Appendix I. The key results are summarized as follows:

We should expect a high number of displaced refugees if:

- A high number of violence is detected in Eastern Equatoria or Central Equatoria and a high number of people are fleeing to DRC or Uganda.
- A high number of violence against civilians are detected and a high number of people are fleeing to DRC or Uganda.

3 RESULTS AND DISCUSSION

During our modelling process, we tried out many algorithms to look for the best results. However, most prediction models didn't perform well with the test sets. This might be due to several reasons: (1) Our interpretation of the data is wrong, we believe that an expert on South Sudan situation may interpret the data differently from us, (2) The data itself should be trained with other algorithms that we didn't try, (3) There are still many factors to consider to predict a better model, (4) We need more data to train a better model. In this section, we will try to summarise all the findings and considerations of this project by answering the two questions posed at the beginning of the paper:

1. How can we use machine learning to assist in predicting human displacement?

Throughout the whole project, we have been using machine learning algorithms to predict and identify characteristics of displaced refugees. An advantage of using machine learning to predict human displacement is that we do not know the main factors that cause these displacements. Use of machine learning might not point out the main factors directly but could help us identify patterns and use those patterns to predict future movements of refugees.

A drawback for using machine learning in this case is that the collected data must be sufficient and accurate, and the accuracy of the model should also be evaluated since providing wrong or inaccurate data to governments or the UN could lead to greater problems with their resource management.

2. What are the main factors that caused displacement of the refugees?

- The increase of violence against civilians is one big factor that forced people to move out from their homes. When there is a high amount of violence against civilians, we can expect a high number of people to move.
- The number of strategic development normally increases during peaceful periods. These kinds of events could easily become the trigger point of the next war.
- The price of goods fell when people were moving constantly; it grew significantly when people returned to the country.
- As is typical, the war started in politically central areas, which are in the southern part of the country of South Sudan. The battle began to move towards the north. Data indicate that when the war started, people from the south mostly migrated to Uganda as it is the nearest neighbor. At the same time, countries to the north, such as Kenya and Sudan, needed to be prepared to receive a mass amount of refugees since people were likely to move away from the north after the beginning of the war.
- From the apriori algorithm, we learned that if both the numbers of "violent events against civilians" and "refugees migrated to Uganda or DRC" were high, the number of refugees moving out of South Sudan would also be high.

4 CONCLUSION AND RECOMMENDATIONS

Our results were limited as it requires the consideration of many more factors to predict the migration flow of refugees during wartime. Using data mining to predict human displacement can be frustrating. The biggest problem we encountered was the quantity and quality of the provided data; most of the time we saw data with many missing values, and it was hard to find additional supporting data for our models. As a result, 80% of our work involved cleaning the data to ensure that the data was predictable.

Summarising from the results, we would like to re-emphasise that the main factor that causes human displacement is hard to judge. A general answer for this is war. From our analysis, we found some minor factors that could influence people, but for accurate results, more data should be added and evaluated (e.g. Educational Status of the South Sudanese, Health Condition, etc). In addition, different machine-learning methods could be tested to find the best results. We have seen researchers uses time-series modelling to predict growth of refugees in Kenya successfully [8], which inspired us to generate the time-series model that we describe in Appendix F. We believe it is worth different methods with the data from South Sudan to find the most accurate results and help the UNHCR to allocate appropriate resources.

APPENDIX

A. CRISP-DM process of data mining

The CRISP-DM model break data mining process into 6 stage, it starts from business understanding or case understanding. Afterwards is working on the data side, from understanding it to prepare it, which includes cleaning the data and make sure that it could be modelled. The process ends with an evaluation to check if the results answer the case's goal. Finally, the results can be deployed if the results is accurate enough [9].

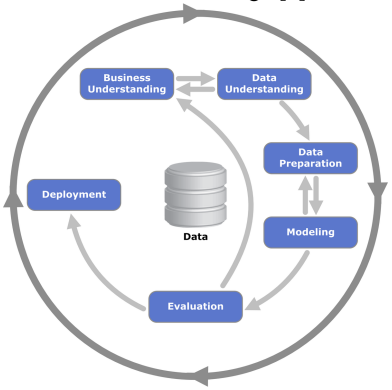

Figure 7 The process of Crisp-DM

B. Meanings of different event types

The meaning of each kind of event types can be seen below in Table 1. The source of this if from the ACLED_Codebook [10].

Table 1 Meanings of different event type

Event Type	Event Description
Battle-No change of territory	A battle between two violent armed groups where control of the contested location does not change. This is the correct event type if the government controls an area, fights with rebels and wins; if rebels control a location and maintain control after fighting with government forces; or if two militia groups are fighting. These battles are the most common activity and take place across a range of actors, including rebels, militias, and government forces, communal groups.
Battle-Non-state actor overtakes territory	A battle where non-state actors win control of location. If, after fighting with another force, a non-state group acquires control, or if two non-state groups fight and the group that did not begin with control acquires it, this is the correct code. There are few cases where opposition groups other than rebels acquire territory.
Battle-Government regains territory	A battle in which the government regains control of a location. This event type is used solely for government re-acquisition of control. A small number of events of this type include militias operating on behalf of the government to regain territory outside of areas of a government's direct control (for example, proxy militias in Somalia which hold territory independently but are allied with the Federal Government).
Headquarters or base established	A non-state group establishes a base or headquarters. This event is non-violent, and coded when a permanent or semi-permanent base is established. There are few if any cases where opposition groups other than rebels acquire territory. These events are coded as one-sided events without a second actor involved.
Strategic development	This event records activity by rebel groups/militia/governments that does not involve active fighting but is within the context of the war/dispute. For example: recruitment drives, incursions or rallies qualify for inclusion. It also records the location and date of peace talks and arrests of high-ranking officials. The inclusion of such events is limited, as its purpose is to capture pivotal events within campaigns of political violence. The notes column contains information on the specifics of the event.
Riots/Protests	A protest describes a non-violent, group public demonstration, often against a government institution. Rioting is a violent form of

	demonstration. These can be coded as one-sided events. All rioters and protesters are noted by generic terms (e.g. Protester (Country)), but if representing a group, the name of that group is recorded in the 'ally' column.
Violence against civilians	Violence against civilians occurs when any armed/violent group attacks civilians. By definition, civilians are unarmed and not engaged in political violence, Rebels, governments, militias, rioters can all commit violence against civilians.
Non-violent transfer of territory	This event describes situations in which rebels or governments acquire control of a location without engaging in a violent act.
Remote violence	Remote violence refers to events in which the tool for engaging in conflict did not require the physical presence of the perpetrator. Remote violence notes that the main characteristic of an event is that a spatially removed group determines the time, place and victims of the attack. These include bombings, IED attacks, mortar and missile attacks, etc. Remote violence can be waged on both armed agents (e.g. an active rebel group; a military garrison) and civilians (e.g. a roadside bombing).

C. Data visualisation of South Sudanese in neighborhood countries

As seen in figure 8, most South Sudanese are displaced in Sudan and Uganda at the moment, The growth of Uganda in 2016 and Sudan in 2017 the relapse of conflicts in South Sudan.

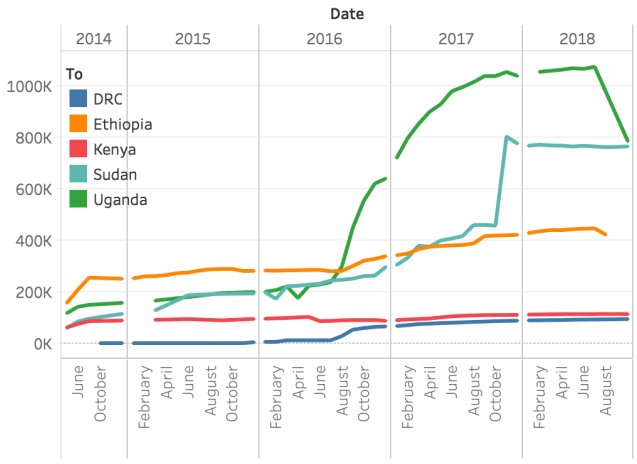

Figure 8 Trend of South Refugees in Foreign Countries

D. Data visualisation of the price of different categories

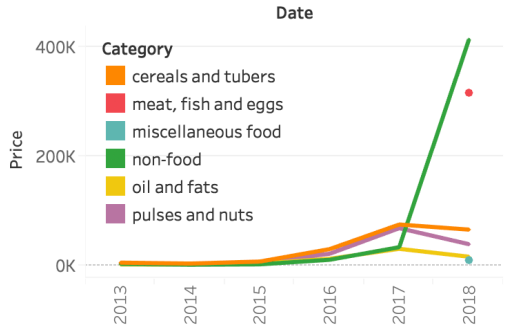

Figure 9 Price of different daily essentials in South Sudan

E. Final Table Structure

Table 2 contains the column names, structure and type of our final training data.

Table 2 The columns contained in our final training model

Attribute	Type	Attribute	Type
Year	int64	Non-violent transfer of territory	float64
Month	int64	Remote violence	float64
Central Equatoria	float64	Riots/Protests	float64
Eastern Equatoria	float64	Strategic development	float64
Gogrial	float64	Violence against civilians	float64
Jonglei	float64	total_events_y	float64
Lakes	float64	total_refugees	float64
Northern Bahr el Ghazal	float64	refugees_move_out	float64
Unity	float64	out_to_DRC	float64
Upper Nile	float64	out_to_Ethiopia	float64
Warrap	float64	out_to_Kenya	float64
Western Bahr el Ghazal	float64	out_to_Sudan	float64
Western Equatoria	float64	out_to_uganda	float64
Battle-Government regains territory	float64	cereals and tubers	float64
Battle-No change of territory	float64	non-food	float64
Battle-Non-state actor overtakes territory	float64	oil and fats	float64
Headquarters or base established	float64	pulses and nuts	float64

F. A Time-Series Model to Forecast Number of Displaced South Sudanese

A simple time-series model was also generated to predict a future number of displaced South Sudanese. The predicted model is presented in Figure 11. We will also like to point out that this is an experience step as we do understand that it is almost impossible to find a trend or sequence for a war. However, it could be possible to see patterns within a smaller period, we will suggest working with experts to look for some sequences as they could be more sensitive with the predicted data, and could include data such as weather or daily essential prices to observe the sequence with the number of people is displaced.

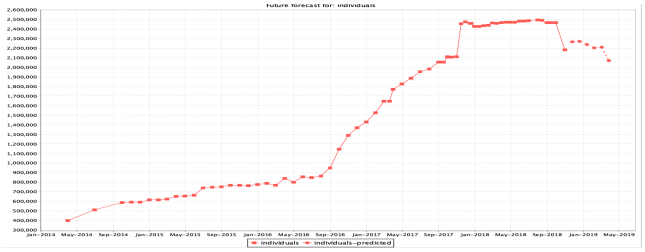

Figure 10 Forecasting the displaces South Sudanese population

G. K-means algorithms

The algorithms for K-means is described in figure 10.

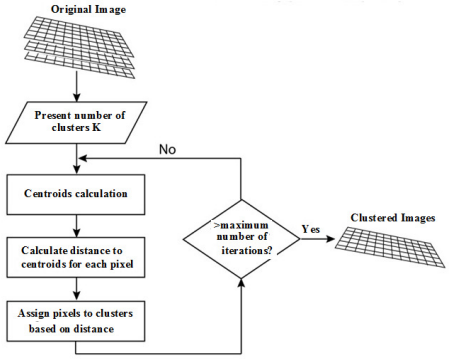

Figure 11 An illustration of the K-Means Algorithm (Figure adapted from figure 12 in Ref. [11])

H. Decision tree to predict fuel prices based on event areas

We found the fuel price of South Sudan were affected by the number of violent/non-violent events that happen in some particular region. The province that are related to the fuel price are Central Equatoria, Eastern Equatoria, Unity and Lakes. The reason for this might be the number of oilfields in that province.

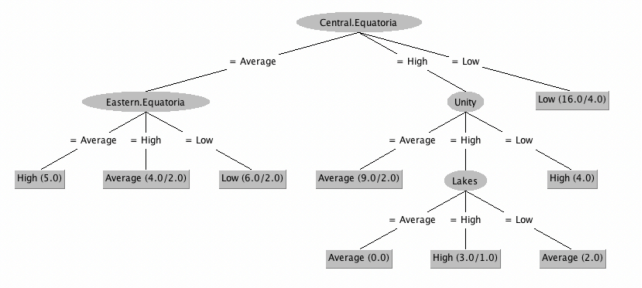

Figure 12 A result from running decision tree for predicting fuel prices

I. Results from apriori algorithm

Item set	Support
Central Equatoria = High, out_to_DRC = High	Refugees_move_out = High
Violence against civilians = High, out_to_DRC = High	Refugees_move_out = High
Violence against civilians = High, out_to_uganda = High	Refugees_move_out = High
Eastern Equatoria = High, out_to_DRC = High, out_to_uganda = High	Refugees_move_out = High
Battle No change of territory = High, Strategic development = High	Refugees_move_out = High
Strategic development = High, refugees_move_out = High	out_to_DRC = High
Central Equatoria = High, Eastern Equatoria = High, out_to_DRC = High	Refugees_move_out = High
Central Equatoria = High, Violence against civilians = High, out_to_DRC = High	Refugees_move_out = High
Eastern Equatoria = High, Violence agasint civilians = High, out_to_DRC = High	Refugees_move_out = High
Eastern Equatoria = High, Violence agasint civilians = High, out_to_uganda = High	Refugees_move_out = High

ACKNOWLEDGMENTS

The datasets used in this project is provided by the UNHCR. The software used in this project were Tableau, Weka, Python (iPython), R. The full project information can be found in: https://github.com/hbshih/South-Sudan-Refugees-Displacement

REFERENCES

[1] Figures at a Glance: 2019. https://www.unhcr.org/figures-at-a-glance.html. Accessed: 2019- 01- 31.

[2] THE UN Refugee Agency, U. 2019. UNHCR PROJECTED GLOBAL RESETTLEMENT NEEDS. 24th Annual Tripartite Consultations on Resettlement. (2019).

[3] OHCHR | UN human rights experts says international community has an obligation to prevent ethnic cleansing in South Sudan: 2019. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?LangID=E&NewsID=20970. Accessed: 2019- 01- 31.

[4] 383,000: Estimated Death Toll in South Sudan's War: 2018. https://www.nytimes.com/2018/09/26/world/africa/south-sudan-civil-war-deaths.html. Accessed: 2019- 01- 31.

[5] 'Untold devastation' in South Sudan triggers grave humanitarian crisis: 2019. https://www.theguardian.com/global-development/2016/aug/09/untold-devastation-in-south-sudan-triggers-grave-humanitarian-crisis-un. Accessed: 2019- 01- 31.

[6] 2018. Situation Overview: Regional Displacement of South Sudanese. (2018).

[7] Kumar, A. and S. Ingle, Y. 2014. Canopy Clustering: A Review on Pre-Clustering Approach to K-Means Clustering. International Journal of Innovations & Advancement in Computer Science. 3, 5 (2014), 28.

[8] Samuel K, M. 2016. A TIME SERIES FORECASTING APPROACH TO PREDICTION OF REFUGEE POPULATION IN KENYA. (2016).

[9] Wirth, R. 2000. CRISP-DM: Towards a standard process model for data mining. (2000).

[10] ACLED 2017. Armed Conflict Location & Event Data Project (ACLED) Codebook. Version 8, (2017).

[11] Sahu, S. 2015. A Support Vector Machine Binary Classification and Image Segmentation of Remote Sensing Data of Chilika Lagloon. IJRIT International Journal of Research in Information Technology. 3, 5 (2015), 191-204.