MODEL CARD: DRC Foresight Model

The Centre for Humanitarian Data's [Peer Review Framework for Predictive Analytics in Humanitarian Response] aims to create standards and processes for the use of models in our sector. The Framework consists of five steps: model submission, technical review, implementation plan submission, ethical review, and final report. The output of the model submission is a completed Model Card that includes information about the intended use, model development, model evaluation, and operational readiness. For further information, please contact Leonardo Milano: leonardo.milano@un.org.

DRC Foresight Model - July 2020

Model Summary

- Organizations developing the model: Danish Refugee Council and IBM
- Model date: 1 March 2020
- Model version: 2.0
- Model type: Machine learning and bayesian network model to predict total forced displacement from a given country 1-3 years ahead (currently covering Myanmar and Afghanistan)[1]
- License: Apache License 2.0
- Main contacts: Alexander Kjaerum (alexander.kjaerum@drc.ngo), Rahul Nair (rahul.nair@ie.ibm.com)
- Links:
  - Repository: https://github.com/IBM/mixed-migration-forecasting
  - Visualisation dashboard: https://drc-tdp-s2-poc.eu-gb.mybluemix.net/home

1. Intended Use

A. In-scope use cases: what is the actual and potential scope of the model? Describe the situations in which the model output is expected to be reliable.

The model is expected to provide a) a prediction of the total forced displacement from a given country 1 to 3 years ahead; b) what-if capabilities for end users to see impact on displacement if selected drivers change; and c) causal network analysis of the different factors driving displacement.

Depending on the level and function in the organization these two outputs can be used for different purposes:

- Global/HQ level: Prioritizing interventions and strategic investments in countries that are predicted to an increase in displacement, global advocacy towards donors using causality analysis to explain consequences of donor policy shifts.

[1] The foresight model has been recently used to estimate the number of people expected to be displaced in the Sahel region by the COVID-19 crisis. We focused our review on the work done for Myanmar and Afghanistan.

- Regional level: With prediction of displacement trends a holistic, data-driven regional strategy can be developed to respond both in host and origin countries.
- Country level: More data-driven process of developing annual strategic planning, HNO/HRP processes including trends analysis and scenario for displacement development, using causality analysis to tailor programming towards root causes.

B. Out-of-scope use cases: what are the model's limits and constraints?

The model is built on macro-level indicators i.e. national level aggregates. By design it therefore does not directly take into account proximate or micro-level factors impacting displacement, nor mediating factors such as cost of movement, changing border policies, etc. There is therefore a limitation in capturing displacement situations where these elements play a major role.

Further, due to the reliance on national level aggregate data, the model so far performs better in situations with large-scale conflict (Afghanistan) impacting major parts of the country. Due to ethical, data access and accuracy considerations the model does not include relevant information on where displacement is likely to happen within a given country, nor where the displaced are likely to go. This limits the operational relevance of the model.

C. Describe the situations in which the model output may not be reliable?

In situations where displacement is more confined (Myanmar), the model performs less well, as the local changes are not necessarily captured in the national level aggregate data. In these instances, a mitigation strategy is to supplement, where available, with more localized data.

D. Model interpretation: what does the output represent?

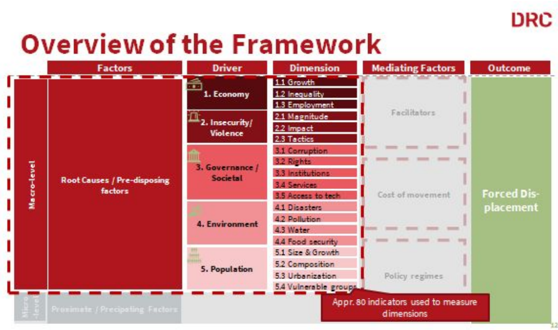
The model outputs are: (a) a point forecast that describes the total stock of forcibly displaced at the country level; (b) a 95% confidence interval of the forecast; and (c) displacement drivers estimated to drive the forecast.

2. Model Development

A. Details of the datasets used to build the model. Describe the sources of data, size and scope of the datasets.

MODEL CARD: DRC Foresight Model

The data is all derived from open source data. The main data sources are the World Bank development indicators, ACLED, UCDP, EMDAT, UN agencies (UNHCR, WFP, FAO), IDMC, etc. In total, the system aggregates data from 18 sources, and contains 148 indicators along several dimensions as shown in the figure to the right.



- Is the data representative of the population being sampled?

Given that the data is taken from reputable data sources, the data is deemed to be representative.

- How accurate or reliable is the training data?

The training data has a few shortcomings. First, coverage is uneven across geographies and across dimensions. For instance, economic and labour statistics tend to have better availability compared to governance and violence statistics. Data from institutional providers can often have a delay. The most recent indicators can be a few years old for many countries.

For training, we limit the data from 1995 through till 2018, the latest data available for displacement. For cross validation, we use a 5-year period between 2010-2015. Following the standard cross-validation setup for time series data, models are trained on data for the years (1995, y) and predictions made for y+t, where y is in the 5-year time period.

One concern is with the historical IDP figures which are derived from the World Refugee Survey up until data is available from IDMC. Another related concern is potential overlaps between UNHCR and IDMC data.

- How is missing data treated? (e.g., exclusion, single imputation, multiple imputation)

The system uses several methods to address data gaps. We distinguish between the missing data in the features (or indicators) and missing target variables (i.e., forced displacement). Data with missing target variables are simply excluded from training. For missing values in indicators, we employ two methods. To address data lag, we make indicator projections for each country using an auto-regressive model (i.e., AR(n) model). An auto-regressive model is a time series forecasting model where future values depend only on previous values of the variable. The 'n' denotes the number of lag variables and is determined using a heuristic approach. For cases where data is insufficient, we simply treat it as missing which is better than projecting incorrectly. Intermediate missing values are computed by interpolation.

Data completeness was investigated for each country and indicator of interest. We generated coverage plots, for example for Afghanistan and Myanmar from 1950. Summaries on missing values are also presented when analysis is run.

B. What are the model assumptions and approximations?

Like most machine learning approaches, the main assumption is that past relationships are valid for the future. By building the models at the national scale, we further assume that displacement is caused by structural factors that are captured by macro-level indicators. Total forced displacement is a combination of internal and external displacement. Caseloads and delayed reporting of asylum cases could muddle the displacement numbers, but the analysis assumes that these issues are not important.

C. Methodology - provide a description of the different analysis steps and how the input datasets are used to train the model.

Data transformations: The first step in the system pipeline is data transformation – the task of standardizing the various inputs so that they can be aligned and used in the model.

Modelling: The machine learning model employed is an ensemble. An ensemble model works by leveraging several constituent models to generate independent forecasts that are then aggregated. Here we employ two gradient boosted trees to generate the point forecasts. The model hyperparameters were determined by means of a grid search. Each year-ahead forecast has a separate model. In other words, we train a set of Ensemble models for $y(t + h) = f(x(t))$, where $h = 0, 1, 2, 3$. The associated confidence intervals were generated by empirical bootstrap method, where the source error distributions were generated on a retrospective analysis.

Scenario "what-if" analysis: The system supports what-if analysis for selected dimensions of economy, conflict, natural environment, governance, and population. For a user-specified scenario, the projections are updated based on estimated elasticities for each cluster. The elasticities capture the change displacement stock for a unit change in the thematic cluster. The elasticities are determined using an ordinary least squares regression. To compute the scenario projections, we use the ensemble model and apply the elasticities from the simpler regression model. For Afghanistan and Myanmar, some of the scenario elasticities are not statistically significant. In these cases, we rely on a basket of 25 countries with a history of displacement to provide an estimate.

3. Model Evaluation

A. Are there other similar existing models and how does this model compare?

No other models have been developed with the same spatial (national level) and temporal (1 to 3 year predictions) focus. A somewhat similar model is the Jetson model by UNHCR.

B. Have there been any past peer reviews, such as published papers or other review exercises?

- “A machine learning approach to scenario analysis and forecasting of mixed migration”. IBM Journal on Research and Development, 23 October 2019
- “Scenario-based XAI for Humanitarian Aid Forecasting”, CHI EA ’20: Conference on Human Factors in Computing Systems, April 2020

C. Is there any reference or benchmark used to evaluate the performance?

For the point forecast, we use a natural benchmark: what happened last year will happen the next. Forced displacement stock between years is strongly correlated so this is a strong benchmark. There is no feasible mechanism to evaluate the confidence interval.

D. What are the metrics used for model evaluation? Why have these metrics been selected? Describe current model performance.

For the point forecast, we considered several error metrics – mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). All these error metrics involve a retrospective analysis, i.e., generating forecasts where the true value was known but hidden. The forecast is then compared with true value to compute the error rate. Since RMSE is quite sensitive to outliers, and it was difficult to compare MAE across different scales of displacement, we elected to do the analysis using MAPE. MAPE shows by how much the forecast is wrong on average and is expressed as a percentage of true value. Lower values are better.

The retrospective evaluation is based on a leave-one-out for time series data. We train the model on a period (start year, end year) and make a prediction for (end year + 1). Specifically, Run 1: training set: (1995 - 2009), Test set: 2010, Run 2: Train: (1995 - 2010), Test: 2011, and so on.

The current model performance, based on forecasts for 1-3 year ahead forecasts for the period 2010 – 2015 are 6.9% MAPE for Afghanistan and 14.6% for Myanmar.

Expanding the test period to cover 2010 to 2019, the MAPE for year 1 forecasts is 8% for Afghanistan and 10% for Myanmar. Compared to the natural benchmark the model performs similar to the natural benchmark (MAPE 11.2%) and outperforms the natural benchmark on Myanmar (MAPE 15.4%).

E. How does performance depend on forecast lead time?

The error rate varies by lead time for the same retrospective validation setup above are:

Lead time (years)	Model MAPE – Afghanistan	Natural benchmark MAPE – Afghanistan	Model MAPE Myanmar	Natural benchmark MAPE – Myanmar
1	6.9%	7.0%	8.1%	15.1%
2	5.7%	4.9%	15.8%	19.0%
3	7.6%	8.0%	20.0%	22.2%

F. What is the risk tolerance of the model? What happens if the model produces false positives, false negatives?

The main risk of the model producing false positives/false negatives are either under over preparedness. The model is not designed to dictate action in isolation but rather inform decision-making along other relevant inputs.

4. Operational Readiness

A. Is the model ready to be used to inform humanitarian response? Is the model kept up-to-date with the latest datasets?

Where possible, the model automatically updates when new data is available.

- **Who is responsible for model updating and/or recalibration?**

The model will be managed by DRC HQ, where a data scientist together with the IT department will be responsible for updating and calibrating the model on an ongoing basis.

- **Is the model ready to be deployed? If not, what are the additional steps needed (e.g., further research, validation, updates)?**

The model is ready to be deployed with only minor work to try to optimize model performance for the two countries.

B. Has the model been developed in collaboration with operational partners? Has the model previously been used in humanitarian situations?

Given that the model has been developed together with DRC, the operational aspect has been integral in the design. Furthermore IBM conducted +10 interviews with various DRC staff at both HQ and field level to better understand humanitarian use cases. The model is yet to be actively used as it has not been finalized to date.

C. What happens if the model is inaccurate, produces false positives or false negatives? Describe expected impact according to the in-scope use cases highlighted in Section 1A above.

As explained above, the model is not designed to dictate action in isolation but rather inform decision-making along other relevant inputs. The main risk of wrong predictions would be failure of a given country strategy and misallocation of resources.

D. List additional considerations for the use of the model in humanitarian response.

The vision is that the humanitarian community will access the model through a user interface. After an initial development phase with IBM, the model will be maintained by DRC and access will open to relevant key stakeholders. The user interface is to act as an analysis platform where users can both access the forecast, develop their own scenarios, but also explore the underlying database of curated data considered relevant for operations. This way the portal will remain relevant even for countries where a forecast is yet to be developed. One potential use case would be HRP and HNO processes where the humanitarian community come together to plan the coming years' response and typically also develop scenarios. Here the model and scenario builder could serve as the interface guiding these discussions.