Modelling IDP displacement: predictors and algorithms

Julian Ibarguen

16/09/2020

Contents

Data protection statement:

The current project have been done with real data. However, the data in its original form is sensitive and private, and its access requires signature of data protection waiver.

To ensure data protection, the original data have been altered: locations names and codes do not correspond to its original; population and displacement figures have been transformed and all data have been aggregated to a superior administrative level; among other alterations in the data and in the report

Despite these modification, care have been taken to keep the structure and proportions of the data so results obtained can be directly extrapolated to the original data. The original data and its processing script can be requested and will be provided prior signature of the standard data protection waiver.

1 Introduction

Following the Internal Displacement Monitoring Centre (2018) there were 30.6 million new displacement associated with conflict and disaster across 143 countries and territories in 2017, either due to natural disaster or conflict. Most of these displacements involve what (UNHCR 2018) defined as Internally Displaced Persons (IDP):

Persons who did not crossed any internationally recognized State borders when they were forced to flee their place of habitual residence, as a result of or in order to avoid armed conflict, generalized violence, violations of human rights or natural or human-made disasters.

The reasons why IDP displace are multiple and diverse, however they are two main events that trigger the massive displacement of IDPs: conflict violence and natural disasters. At the onset of a crisis trigger by these events it is likely that there is massive displacement being the reason of displacement the event it self. Once the crisis starts to settle, and in the worst scenario becoming protracted, the displacement is likely to reduced in quantity and stabilize. Many IDP would further displace to other locations with improved living conditions, other would displace due to stemmed consequences of the main event or progressive deterioration of living condition at their location; other would return to their original locations.

As the crisis becomes protracted, secondary displacement (i.e. IDP that have displaced to two or more locations different than its origin) becomes more predominant. IDPs on secondary displacement face progressively increased vulnerabilities due to the fact that they not only lose everything at origin, they also lose everything at their initial displaced location, accumulating the deterioration of their living condition due to the fact of displacement itself, but also due to duration of the displacement

In this document we aim at exploring the possibility of defining a model to predict IDP displacement using predictors related to the displacement and the living condition at departure and arrival locations, leaving aside predictors related to the events. For this we will use monthly collected data at location (community) level for an interval of 20 months. For each month the unit of analysis was the itinerary (aggregated at municipal level). These itineraries were described with three thematic areas of variables: displacement, demographic characteristics, sector needs and reason of displacement. This thematic areas corresponded to progressively more difficult to acquire and less readily available information. The data was split between training and hold-out. Further over the training set we applied a Rolling Origin Forecasting to create the Probe data

We defined a baseline model with the historical average displacement by itinerary (historical as to the period covered by the study). Aside we defined our final model in three steps incorporating predictors by thematic batch: the first one for displacement thematic predictors, the second one we added demographic thematic predictors, the third included sector needs and in the fourth we added reasons of displacement. We observed the gains in the Root Mean Squares Error (RMSE) against the baseline model and between the three progressive models as to observe the improvement of the model. Each model was processed with three trained algorithms: Least Absolute Shrinkage and Selection Operator (LASSO), Partial Least Squares (PLS) and Random Forest (RF).

At the end of the process the model and algorithm that provided the lowest RMSE was chosen and its predictive capacity tested against the hold-out data. Aside of the predictive model, the step approach for modellign allowed us to observe the performance of the predictors as they are progressively added into to the model. Finally, but not least, we further obtained a valuable exploratory analysis on determinants of internal displacement.

2 Methodology

2.1 Data

The data was monthly collected data of internal displacement at community level from January 2017 to August 2018. It was collected by mean of key informant interviews reporting figures on arrival and referring the most common place of departure for each location of arrival. In addition to internal displacement figures, data on the needs of the population and key reasons for displacement was further recorded.

Needs of the population were classified in eight sectors traditionally associated with humnaitarian action and were reported as percentage of persons in need over the total population for each sector; data on reasons of displacement referred to pull and push factors and were rated in a three level scale: from *"Not important"* to *"Very important"*. Pull factors referred to reasons in arrival that motivated persons to displace to the location; push factors refered to reasons at departure that motivated persons to displace from the location.

Data was aggregated at municipal level and arranged to represent as unit of analysis the itinerary. This is, each row represented a displacement from one municipality of departure and to another municipality of arrival. Each itinerary was attributed with persons displaced, total population at arrival and departure, persons in need at arrival and departure and scores for pull factors and push factors.

All in all, in 20 months the data set compiled a total of 4,289 itineraries, being 50.8% of them traveled more than one month. For that period the total number of IDP was 7,601,435. On average there were 214 itineraries and 380,072 IDP displaced per month.

2.2 Cross-Validation

The data was split in two from the beginning, holding-out of the analysis the last two month available. The hold-out months were used at the end to test the final accuracy of the model, the remaining data was used for training the model. The training data was further split in six folds for an evaluation on what is commonly known as Rolling Forecasting Origin.

Rolling forecasting origin evaluation was defined with a cumulative window starting at 10 month and an horizon on 2 month. This is, the first fold was composed with data from the first 10 months and aimed at predicting the next 2 month (11 and 12), the second fold was composed by data of the first 11 month, and aimed at predicting the following 2 months (12 and 13), and so on up to the sixth fold. Furthermore, each of the fold was bootstrapped for 15 re-samples.

This data partition strategy had several benefits:

1. The usage of Rolling forecasting origin allowed us to apply a Cross-Validation (CV) approach without violating the principle exposed by (Bergmeir, Hyndman, and Koo 2018), by which CV, in its original form, should not be used when model residuals are strongly correlated, as was proved for our case.
2. The CV allowed us to get the maximum benefit of our sample size to reduce over-training and to testing the model's prediction at different point in time.
3. The bootstrap allowed us to further expand the benefits of the CV folds, by resampling each of the folds in up to 15 different samples.
4. Finally, the hold-out set of the last two month allows to test the final model in a non-trained data, using the traditional out-of-sample approach, which is a traditional standard in time-series when residuals are correlated.

Bergmeir, et al. (2018) suggest the Ljung-Box test to asses the adequacy of using traditional CV over time series. The result of the test for our final selected model was significant at 0.006, rejecting the null hypothesis and assuming dependency across the model's residuals. This justified our election of a more complex CV strategy.

2.3 Modelling approach

The main objective of this work was to define a model to predict internal displacement on monthly basis, thus our outcome was the number of IDP. Nonetheless, as we will show later the outcome was log-normal and we used its log transformation for the modeling.

Our predictors were grouped in four main thematic areas: displacement, demographics, sector needs and reasons of displacement. For each of the thematic areas, predictors were explored and new predictors created and transformed as relevant. The defined models were processed with three algorithms: LASSO, PLS and RF. The selection of this three algorithms was motivated by the following criteria:

- Able to resolve regression problems
- Provide indicators for variance importance and good interpretability of results
- Preference for linear methods, given nature of data, but at least one non-linear algorithm
- Balance with computation requirements

A baseline model was defined as a simple Linear Regression with the historical average of IDP displaced per itinerary. Aiming at outperforming this baseline, but also to explore the impact of new predictors on the model's performance, the final model was constructed in four steps: in each step a new batch of predictors corresponding with a new thematic area was added to the model and the algorithm was tune accordingly.

At the end of the fourth step the model and algorithm that provided lowest RMSE was selected and tested against the hold-out data, which was fully excluded from the exploratory analysis and the model training.

3 Determinants of internal displacement

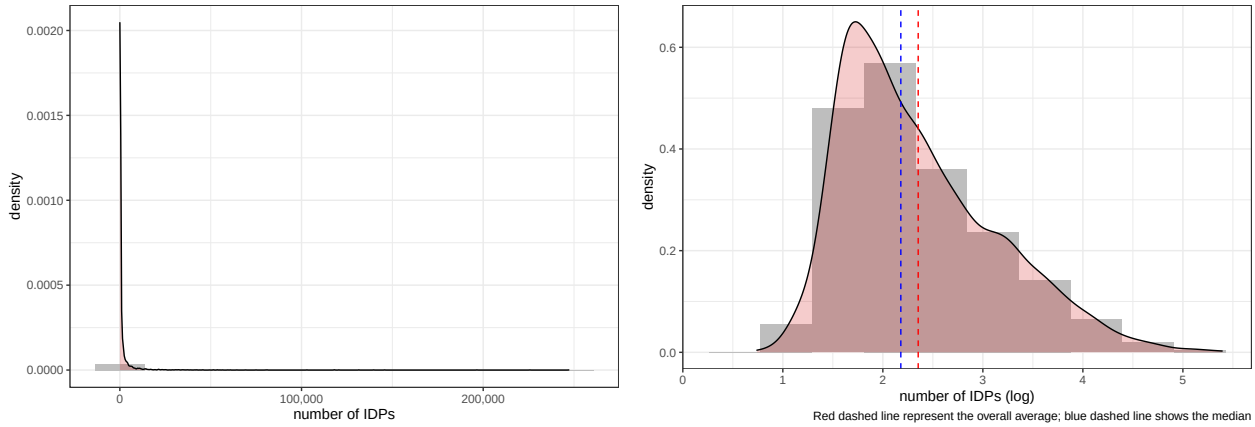
The exploratory analysis was done over the training data, which includes months from January 2017 to June 2018. Despite of having exclude the last two month available for 2018, still the exploratory analysis serve as fair descriptive analysis on determinants of internal displacement

3.1 Displacement

Displacement variables are intrinsically related with the event of displacement it self, such as the time, the distance, the location of departure and arrival. This variables are readily available in any data collected about displacement, as they are the basic information to attribute displacement and need of no external or complementary information.

3.1.1 Number of IDP: the outcome

Before starting with the predictor, first we described the outcome of our model. The distribution of the number of IDP displaced had a median for the study period of 151 with an inter-quantile range of 608, describing an extremely skewed distribution (Figure 1a). The distribution was a log-normal, thus after logarithmic transformation with base 10 it approximate the normal distribution (Figure 1b). The center of the logarithmic distribution was at 2.35 (2.35 standard deviation - SD)



(a) Non-transformed distribution

(b) Log10-transformed distribution

Figure 1: Density distribution of number of IDPs displaced

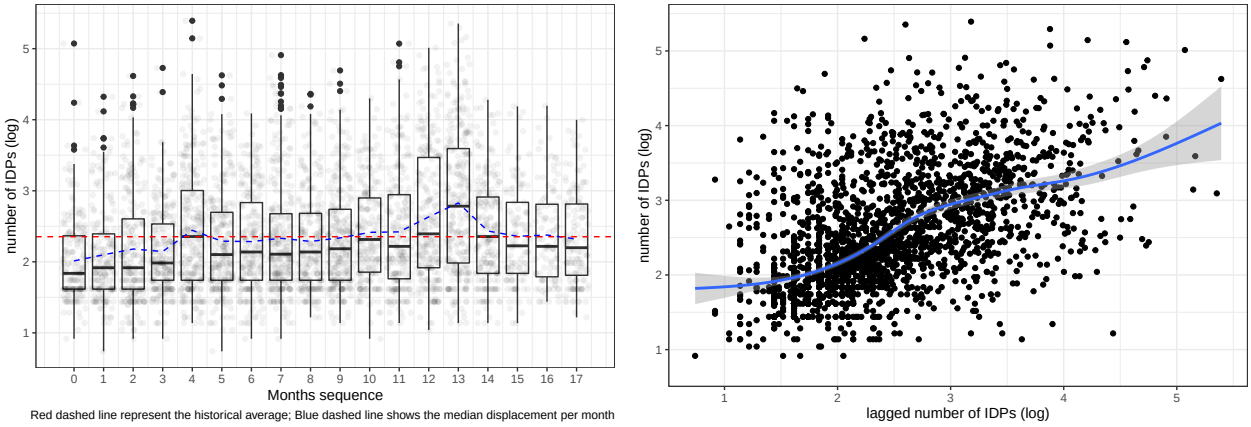
Given the extreme skewness of the outcome and the positive response to logarithmic transformation, we decided to develop our model in its log-transformed state. The logarithmic transformation offered a more normal distribution likely to perform beter in our prediction model, as previously proved by Anonymous (2018). Nonetheless, a more normally distributed outcome came at expenses of limitations on the model's interpretability. Despite this lost on interpretability, priority was given to model performance as the main objective of the study was predictive and not explanatory. Furthermore, the logarithmic data could be transformed back to refers to the original unit in its natural scale.

From now on, the number of IDP displaced was treated in its transformed scale and all graphs and result tables presented below refers its logarithmic transformation.

3.1.2 Time

IDP displacement is mainly influenced by critical and sudden events, which could be somewhat stationary such as conflict and disaster. However, specially on protracted crisis, more contextual factor such as livelihood, food and access to services would start playing a important role on IDP displacement. These new determinants gain importance at the stage of settleling from the crisis and are influence by time, and indirectly IDP displacement might feel the effect of time on these deteminants.

When we observe the distribution of displacement across the study period we appreciate a more or less homogeneous base of displacement across time, pointing ato a protracted crisis scenario with an estable base of displacement likely to be more determine by contextual factors than sudden events. Reflection of this is the overall little variation of displacement from one month to another with a SD of 0.19. Despite of this general estability, we also appreciate peaks with sudden onsets of displacement (Figure 4a). It is likely that these peaks are cause by sudden events of violence and/or natural disaster, accounting for the greater of the variability od IDP displacement through time.



(a) Trend of IDP displacement across study period

(b) Relationship between lagged dispalcement and current displacement

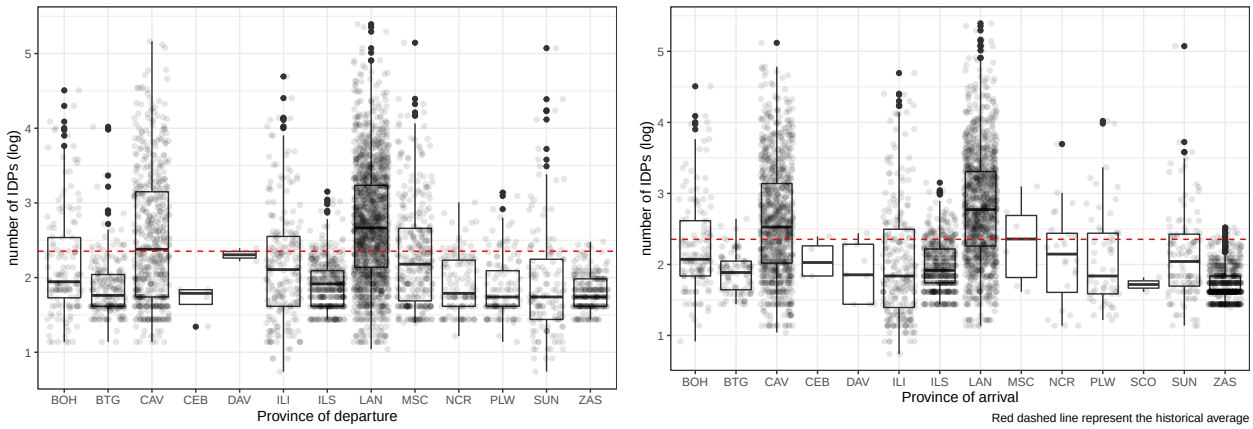
Figure 2: Trend of IDP displacement across study period

Our model needed to have the capacity to capture the stable trend, at the same time that the challenge of capture the sudden peaks in displacement. Factors inherent to dispalcement and more socio-demographic variables are likely to better capture the estable base of displacement (figure 4b - show hoe the dispalcement of the previous month could serve as predictor of the next month, which in cases of high intermonth vairability would not be the case).

In addition, and to complement the performance of our model, Information on critical and sudden events would be key to capture the peaks of displacement, despite this information is rather scarce and not inmediately available.

3.1.3 Itineraries

Regarding the influence of the itineraries in the displacement of IDPs, in Figure 3 we can observe that mainly two provinces (LAN and CAV) accumulate a great part of the displacements, both for the departures and the arrivals. Other provinces are more likely to be provinces of arrival than of departure, such as ZAS, CAV and ILS, reflection of inter province displacement. When we observe the SD of the province averages it is 0.31 for both, provinces of departure and arrival. These diversity in the displacement at the province level is likely to further powered at municipal level and would be pointing at the effect of the itinerary in the displacement of IDP.



(a) Province of departure

(b) Province of arrival

Figure 3: Distribution of IDPs displaced (log) by province

To further explore on the effect of itineraries over displacement, we calculate the Historical average of displacement by itinerary and plot it against the number of IDPs displaced (Figure 4). Not by surprise we can clearly observe a linear relationship between the itineraries' average and the displacement of IDPs. Seeing the strong relationship between both and the simplicity of the predictor, Historical average of displacement by itinerary was choosen as the the predictor to our baseline model, the one we wanted to outperform.

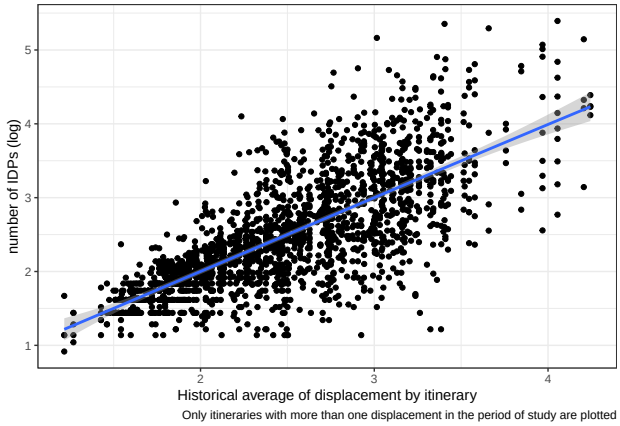
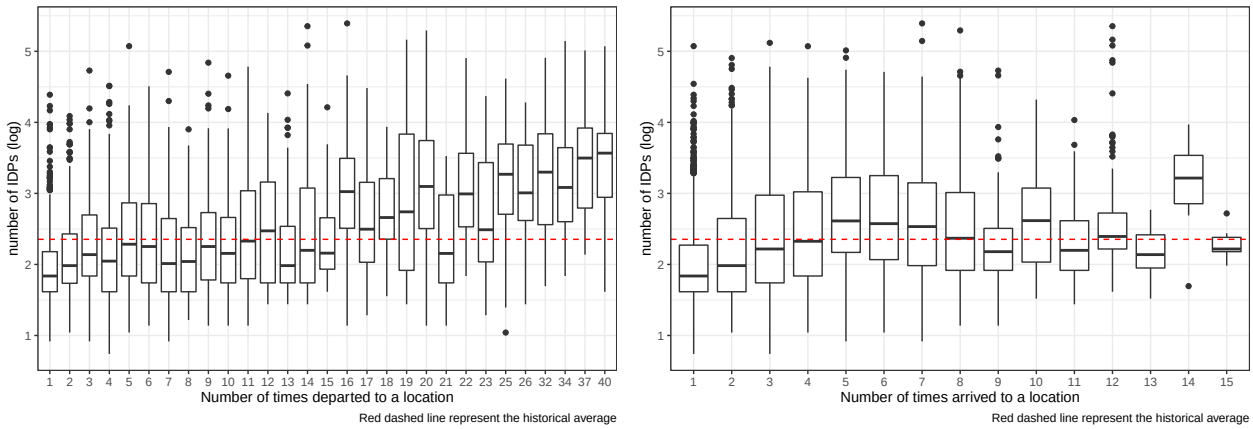


Figure 4: Itineraries' historical average displacement by number of IDPs displaced

3.1.4 Number of times municipality of departure/arrival

Another important element of the itinerary, is how many times displacement have departed or arrived to a municipality. Observing Figure 5 we could see that that number of times departed from a municipality shown a linear trend with IDP displacement and with relevant variability (SD = 0.43). However, this was not as clear for number of time arrived to a municipality which follow a more stable trend (SD = 0.28), still it shown centain vairability that coudl be useful to predict certain scenerarios.

Meanwhile the historical average displacement served as measure of density of displacement from the location; number of times departed/arrived to location acted as a measure of frequency of displacement. The two measure complement each other to further attribute IDP displacement and explain its variability.



(a) Number of time departed from municipality

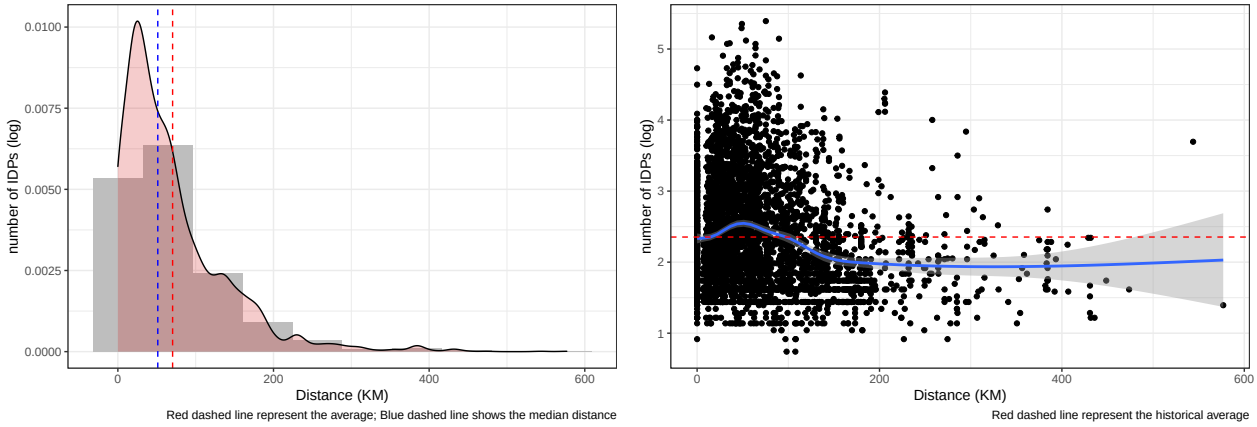
(b) Number of time arrived to municipality

Figure 5: Distribution of IDPs displaced (log) by number of times departed/arrives to a municipality

3.1.5 Distance

The itinerary is also characterized for the distance that comprise between the departure location to the arrival. The distance of the itinerary was calculated by estimating the geodesic distance from the municipal centroids corresponding to each itinerary. Although distance was also a log-normal variable skewed to the right, we studied it in its original form.

A priory would be expected that IDPs displace preferentially shorter distances, afollowing a negative linear relationship between distance and displacement. Nonetheless, when we observe the relationship we see it is not a clear linear relationship. Although generally larger distances imply lower loads of displacement, the shortest distances also imply lower numbers of IDP displacement.



(a) Distribution of distances

(b) Relationship between distance and IDP dispalcement (log)

Figure 6: Relationship between distance and displacement

An explanation of this might be the level of analysis wer are taking. Might be the municipality is a small area for which the shorter distances does not really reflect a change in the conditions. In other words, an IDPs would not move to a nearby location that is in the same situation as its current location.

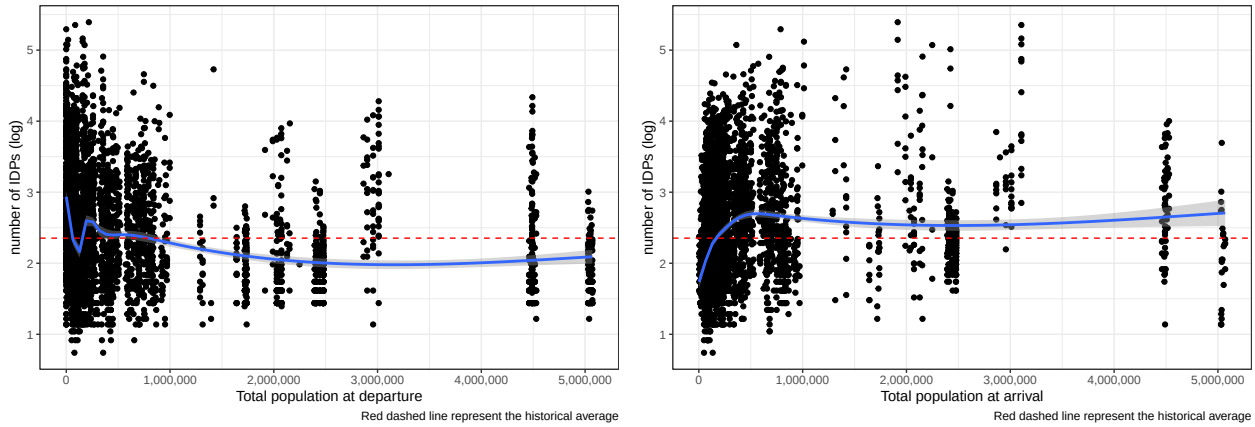
3.2 Demography

Demographic factors are related to the composition of the population in the locations of departure and arrival. Here we consider total population, IDP rates and rates of displacement and arrival. Although rates of displacement and arrival could well have been classified as displacement variables, they need total population data to be created and thus not fully inherent to displacement data.

Meanwhile displacement thematic area contains information inherent to any displacement data collection, demography thematic area usually require additional questions or complementary sources of information. Although, this information is usually available it is frequently not updated or with important lags.

3.2.1 Total population

Starting with the total population, we observe (Figure 7) that meanwhile seems IDP tend to displace from municipalities with low total population (ρ = -0.2), it does seems that people tend to displace and arrive to more populated municipalities (ρ = 0.17).



(a) Total population at departure

(b) Total population at arrival

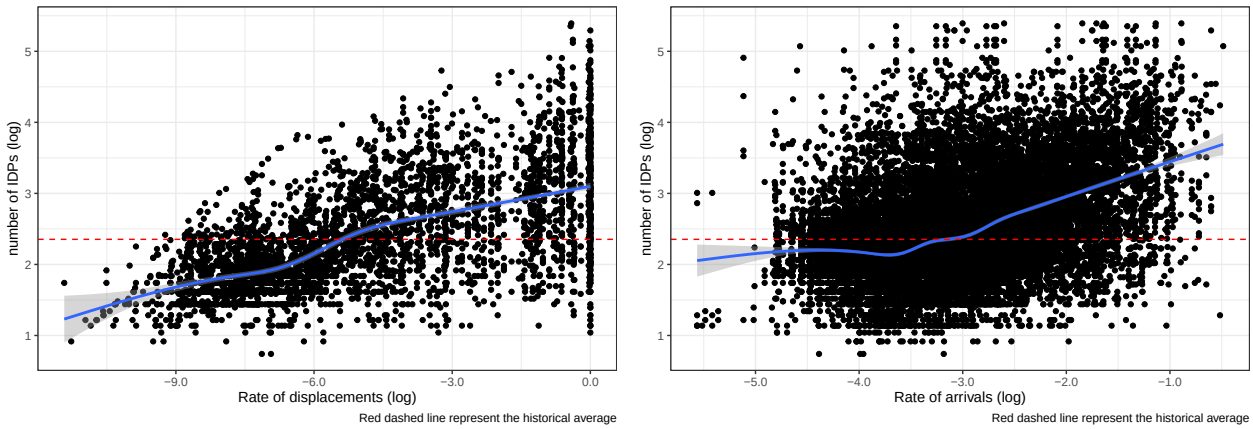
Figure 7: Relationship between total population and displacement

3.2.2 Rates of displacement and arrival

The relationship between total population at arrival and at departure, and the number of IDPs displaced can be drawn by the rates of displacement and arrival. This rates are defined as the proportion of IDPs displaced relative to the total population at departure and at arrival location. Similar to the historical average of displacement dis with the itinerary, these indicator have the advantage that directly synthesize and put in relationship the predictor, for this case total population, and the outcome.

Both rates, displacement and arrival, where log-normal and as in nature related directly to IDP displacement, they were also log-transformed for its inclusion in the model. For the rate of displacement, the total population of the previous month for a municipality of departure was used, to account for the time lag stemmed from the fact that interviews were recorded at arrival, after the departure. To impute the first month after the lagging, we kept its own value. This was preferred over other imputation methods such as median or other central measure as it kept intact potential growing or decreasing trends from the first month of study.

In Figure 8 we can observe, as expected, that both rates relate linearly to IDP displacement. However this relationship seems stronger in rate of displacement than in rate of arrival, with a correlation coefficient (ρ) of 0.43. The coefficient for rate of displacement was of 0.63



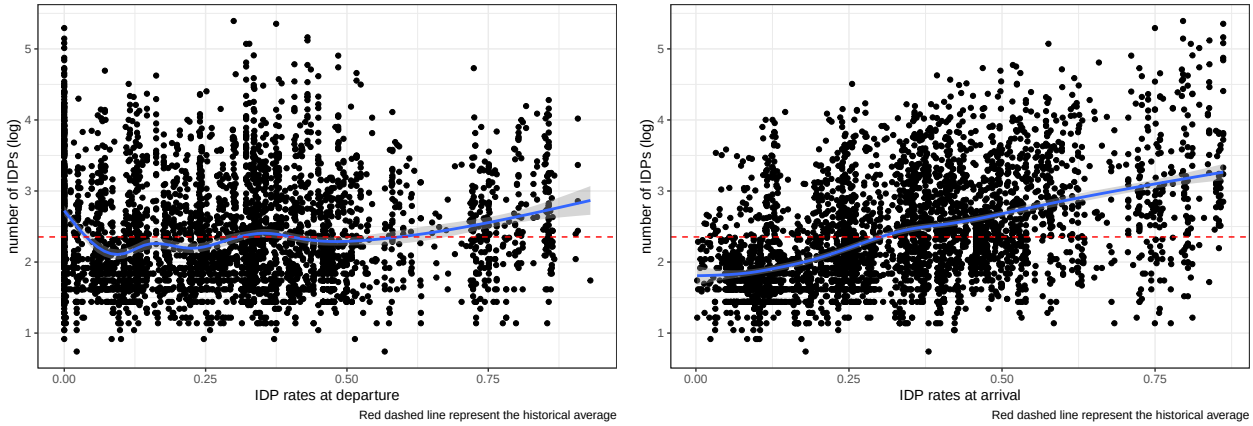
(a) Rate of displacement

(b) Rate of arrival

Figure 8: Rates of displacement and arrival, and IDP displacement

3.2.3 IDP rates

Another measure of interest related to demographic thematic area was the proportion of IDPs in a specific municipality of departure and arrival. The creation of this indicator is totally independent of IDP displacement and combine two population variables: total population and IDP population, forming what we could call IDP rates. From Figure 9 we can see that meanwhile IDP rate at departure have little effect on displacement ($\rho = 0.04$), IDP rates at arrival show a strong relationship ($\rho = 0.5$).



(a) IDP at departure

(b) IDP rates ar arrival

Figure 9: IDP rates and displacement

This might be explained by the fact that IDPs tend to group in camps at their arrival location, thus location with IDP camps are more likely to attract IDPs and thus displacement is higher in municipalities with higher IDP rates.

3.3 Sectorial needs

Information on sector needs is frequently collected on rapid assessment and can be extracted and composed from big household surveys. However the first is rarely available for the whole country, sparse and limited to

specific areas; meanwhile the latter is not frequent and demand high analytic capacities to build composites to extract sector needs indicators.

The collected data considered here cover 8 sectors: health, education, basic services, livelihood, water, security, food and non-food items (NFI). For each of the sectors the data represent the percentage of the total population in need of an specific sector. Sector needs are considered at location of departure and location of arrival. This form of collecting sector needs corresponds more to rapid assessments.

In Table 1 one we can observe the correlation coefficients (Pearson) for the relationship between each of the sector needs and IDP displacement. At departure the highest coefficients we found them in the needs of water and food, similarly for arrival are food, NFI and water the needs that more correlates with displacement.

Although overall all needs correlation with IDP displacement, education, basic services and livelihood at departure have a very weak correlation; livelihood is also weakly correlated in arrival. One explanation of why overall arrival needs are more correlated to displacement that departure needs is because IDP in the displacement have lost all they have before and it is in arrival when their needs increase. Thus needs in arrival are a consequence of the displacement, while need in departure are triggers of displacement.

Table 1: Pearson correlation coeficients between sector needs and IDP displacement

sector.need	**at.departure**	**at.arrival**
need.health	0.169	0.261
need.water	0.198	0.371
need.food	0.192	0.401
need.nfi	0.144	0.391
need.education	0.093	0.236
need.livlihood	-0.058	-0.033
need.security	0.168	0.291
need.basicservices	0.084	0.324

When observing the correlation structure between sector needs at arrival and departure, we observe that effectively most of the sector needs are correlated between them for each place of the itinerary (arrival and departure). However, need at departure and needs on arrival are weakly correlated. Exception is livelihood needs weakly correlated at arrival.

All these suggest that sector needs are susceptible of factorization and would be better included in the model after dimensionality reduction through Principal Component Analysis (PCA) for each group of needs, at arrival and at departure. Factorization allowed us to develop a more effcient model reducing the number of predictors, btu also to better account for the correlation between the sector needs/ In exchange, we lost interpretability of the model. We will show the dimensionality reduction when introducing this batch of indicators into the model: at Model 3.

3.4 Reasons of displacement

In terms of data collection, reasons of displacement follows a similar approach than sector need. However, they are more rarely collected in rapid assessment and they are likely to be the most difficult type of data to achieve, specially at very disaggregated level as the one we are addressing here: the municipality.

In this study Reasons for displacement referred to a series of pull and push factors. Push factors were conditions at departure that could motivate the displacement from the location. This were rated in a three level scale: *No important*, *Important* and *Very important* as to refer the degree of importance in determining the displacement.

When we observe at how the different factors correlate with displacement (spearman coefficients) we observe the most determinant is security, either improved security at arrival or worsen security at departure; followed

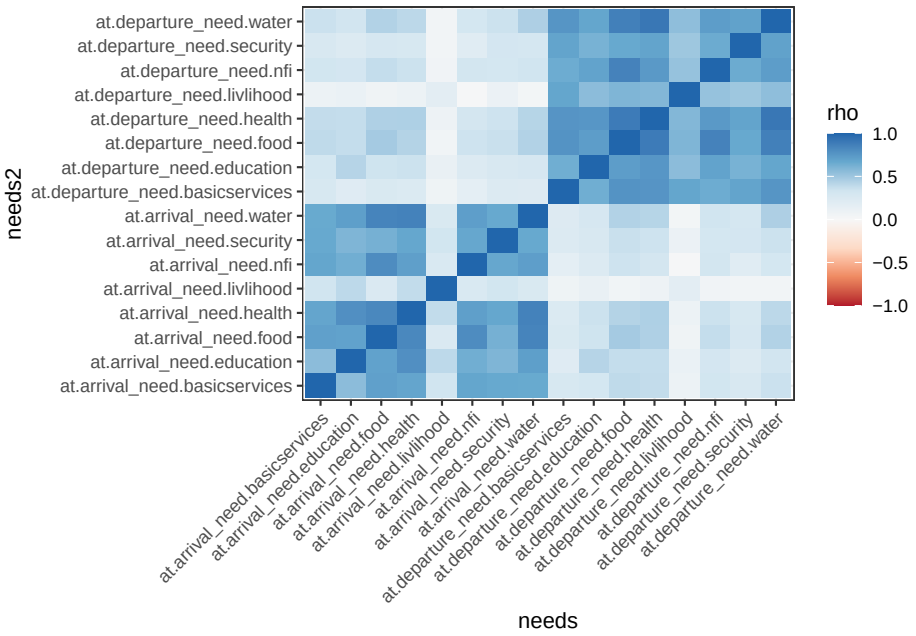


Figure 10: Corrrelation structure of sector needs at arrival and departure

by services, either availability of services at arrival or lack of services at departure. Also worth noting the fair correlation between access to humanitarian assistance at arrival and IDP displacement

Economy showed a negative relationship with displacement being that worse economy at departure and improved economy at arrival associated with lower IDP displacement. Meanwhile improved economy at arrival could be a reasons discouraging IDPs to do displace further, it more difficult to understand why worse economy at departure do not motivate displacement.

Table 2: Spearman correlation coeficients between pull/push factors and IDP displacement

factors	**push**
WorseEconomy	-0.387
WorseSecurity	0.396
LackServices	0.265
Evacuation	0.092
Other	-0.031

factors	**pull**
ImproveEconomy	-0.380
ImproveSecurity	0.410
AvailServices	0.222
AccHumanitarian	0.302
SocialCulturIssues	0.114
InTransit	0.120
NoOption	0.088
Other	-0.034

Unlike for sector needs, the correlation structure for pull and push factors is overall weak and diverse. We discarded conducting any dimensionality reduction analysis over this variables and rather consider each of them as predictor.

4 Modelling

After the exploration of the different independent variables and its relationship with number of IDP displaced, new predictors were identified, These predictors were added to the model together with the original independent variables: 1) Historical average displacement by itinerary; 2) Number of times municipality of departure/arrival; 4) Rate of displacement/arrival and 5) IDP rates at departure/arrival

In this section we will present the different models that we considered for this study. Model were processed by three algorithms and draw progressively in four steps including in each a new batch of thematic predictors. Model performance was evaluated with RMSE, which have the benefit of reporting the error at the same unit of analysis. Nonetheless, R2 and Mean Absolute Error (MAE) were also reported.

The models were processed with three algorithms:

1. **LASSO** is a penalized linear regression method that minimize the sum of squares through a penalizing term refered as tuned by a λ parameter ($\hat{\beta}_{(\lambda)}$. For predicting the outcome $\hat{y_i}$ with predictors $x_1, x_2, x_3 ... x_n$ The model can be notated as:

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta}_{1(\lambda)} x_1 + \hat{\beta}_{1(\lambda)} x_2 ... \hat{\beta}_{n(\lambda)} x_n + \epsilon_i$$

2. **PLS** is a regression model which build on similar pricniples of dimensionality reduction as PCA. Unlike PCA, the PLS method decomposes both, the matrix of predictors and the outcome as a common set of specific factors that explain as much as possible covariance between both. In a later step, the matrix of factors is regressed into the outcome. Thus model for prediction could be notated as:

Where a matrix of predictors X is decomposed $X_{(nxp)} = TP^T$, being T the score matrix and P the loadings matrix; and the outcome Y is decomposed in $Y = UQ^T$. Then X is regressed to Y with:

$$\hat{y_i} = T_i Q_i + E_i$$

3. **RF** is a method that ensemble multiple decision trees to gain accuracy. Trees grown by method of bootstrapping the sample and splitting the data set by nodes selected according to the values of a randomly selected subset of predictors. For our study we decided to grow 250 trees for each model. Once grown the trees, the "forest" is ensemble to form the prediction. The ensemble stage of the RF method could be notated as:

$$\hat{y_i} = \frac{1}{T} \sum_{t=1}^{T} \hat{y_i}$$

Where T is a decision trees in the RF

4.1 Baseline: Historical average

Historical average shown a strong linear relationship with IDP displacement, further its calculation is simple and the information is readily available with any data collection on IDP displacement. This was the reasons it was used as the baseline model, which our final model would outperform. The baseline model was defined for each itinerary (i) as the linear regression between the outcome ($\hat{y_i}$) and the predictor$\bar{y_i}$. The model was notated as:

$$\hat{y_i} = \beta_0 + \beta_1 \bar{y_i} + \epsilon_i$$

After tuning the model the best β was found at 1 The RMSE of the model was 0.4336, which was a fair RMSE for a baseline model. This proved the linear relationship between bothg, otucome and predictor, indicating historical IDP displacement average is a fair predictor of IDP displacement by itself

Table 3: Baseline model evaluation: Historical average of the itinerary

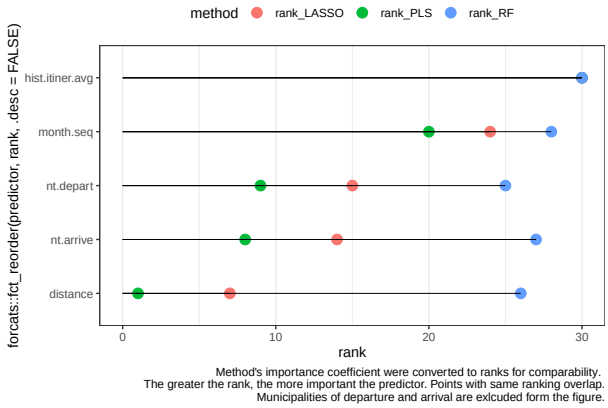
model	RMSE	Rsquared	MAE
Historic Average	0.4336	0.7261	0.2829

4.2 Model 1: Displacement

In our first step we added to the baseline model the predictors identified from the displacement thematic area, this were:

- Month sequence: time of displacement (month and year);
- Distance: distance in meters between centroid of departure municipality and of arrival municipality,
- Municipality of departure: municipality from where the IDPs departed
- Number of times municipality of departure: how many times the municipality was the place of departure IDPs during the study period
- Municipality of arrival: municipality where the IDPs arrived
- Number of times municipality of departure: how many times the municipality was the place of arrival IDPs

After processing the model with the three defined algorithms and tuning each of them we observed that none of the models and method was able to beat the baseline model (Figure 11). Among them, LASSO and PLS, linear regression methods in nature, performed significantly better than RF. Regarding predictor importance Historical average of displacement by itinerary was the most important for all methods, followed by Month sequence. Distance was overall regarded as the least important predictor by all methods, excluding Municipalities of departure and of arrival.



(a) Predictor importance ranking

model	RMSE	Rsquared	MAE
LASSO	0.4398	0.7184	0.3121
PLS	0.4400	0.7183	0.3124
RF	0.4697	0.6846	0.3226

(b) Model evaluation parameters

Figure 11: Model 1 performance

The best RMSE was of 0.4398 with LASSO method. The most important predictor, as agreed by all methods was Historical average of displacement by itinerary.

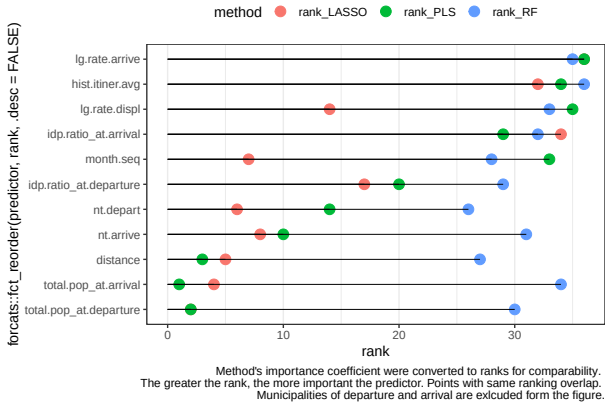
4.3 Model 2: Displacement & Demography

Following our approach, in the second step we added to our model the demographic predictors, which were defined as follows:

- Total population at departure
- Rate of disaplcement (log): the proportion of IDP displaced over the total population at municipality of departure
- IDP rate at departure: proportion of IDPs in the location of departure
- Total population at arrival
- Rate of arrival (log): the proportion of IDP arrived over the total population at municipality of arrival
- IDP rate at arrival: proportion of IDPs in the location of arrival

After processing the model through the three identified methods we observed a significant improvement in the prediction capacity of the model. The baseline model becomes clearly outperformed and RF takes over to be the best performing method (Figure 12).

Meanwhile Historical average of displacement continue to be on of the most important predictors, Rate of arrivals (log) and IDP rate at arrival are new predictor which have gained relevant importance across the three methods.



(a) Predictor importance ranking

model	RMSE	Rsquared	MAE
LASSO	0.2053	0.9329	0.1614
PLS	0.2048	0.9334	0.1603
RF	0.0838	0.9894	0.0452

(b) Model evaluation parameters

Figure 12: Model 2 performance

The best RMSE was of 0.0838 with RF method. The most important predictor, as agreed by all methods was Rate of arrivals (log).

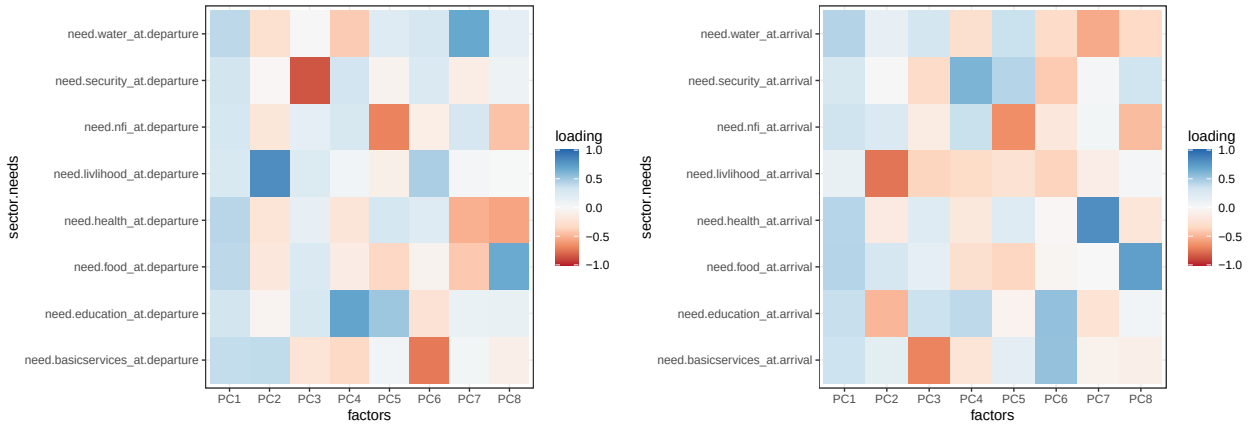
4.4 Model 3: Displacement, Demopgraphy & Sectorial Needs

At the third model we entered the sector needs predictors. As we mentioned before during the exploratory analysis, these set of variables was susceptible of PCA. We conducted the analysis over the five folds of the training data. We targeted to keep the mode number of components among the five folds that would account for at least 90% of the variance. At the end we kept four components.

PCA came at expenses of lost of interpretability fo the model, however the loading of the component would help to better understand what each component represent. This way, in Figure 13 we plotted the loading in a heat map for the component of sector needs at departure and arrival. The darker the blue, the strongest is the weight of a original variable over the component (i.e. the more relevant is the variable to explain the component). Same, but in inverse direction (i.e. negative relationship), for dark red.

As we could see in Figure 13a, the first four components for sector needs at departure municipalities represented: 1) the first component represent those municipalities with average number of persons in needs across all sectors; 2) the second one identify those municipalities with greater number of person in need of livelihood; 3) the third component outstood for those municipalities with significantly reduce number of persons in need of security; 4) the fourth and last chosen component referred to those municipalities with greater number of persons in need of education.

In Figure 13b we can do similar analysis for the sector needs at arrival, noticing that 1) the first component, similarly to departure, represent those municipalities with average number of persons in needs across all sectors; 2) the second highlights those municipalities with less number of persons in need of livelihood, and up to a less extend education; 3) the third refer to municipalities with low number of persons in need of basic services; 4) the fourth and last selected component up to less extent referred to municipalities with greater number of persons in need of security.



(a) Components loadings at departure

(b) Components loadings at arrival

Figure 13: Principal Components loadings for sector needs predictors

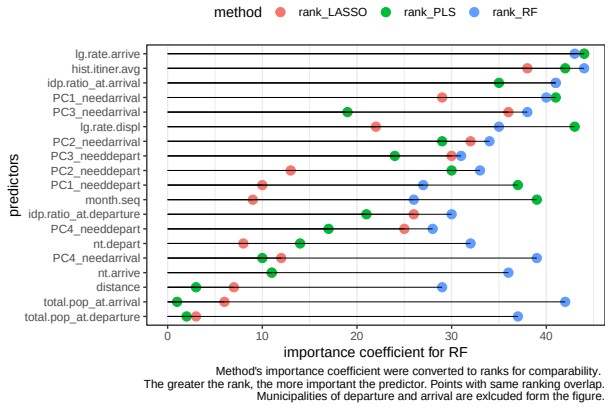
After running the three algorithms over our model we observe a slight improvement on performance in LASSO and PLS, and a slight worsening RF. Although changes are very minimal, at these levels of accuracy even the slightest changes are significant. The RF method still performs better than LASSO and PLS consolidating its predominance. Regarding the new predictors, we can observe that three out of four of the component have entered in the top 10 more important predictors, thsi is true also for the RF method.

The best RMSE was of 0.0893 with RF method. Overall, most important predictor, was still Rate of arrivals (log), which performed best fro LASO and LSP method, and second best fro RF. Historical average displacement of the itinerary was the best for performing predictor for RF

4.5 Model 4: Dispalcement Demography, Sectorial Needs and Reasons of Dispalcement

Finally, we reach our forth model in which we included the reasons of displacement (a.k.a pull and push factors). All push and pull factor were rated to reflect the degree of importance in determining the displacement. The following were included

- Push factor - Worsen economy, at departure location
- Push factor - Worsen security, at departure
- Push factor - Lack of services, at departure
- Push factor - Evacuation, forced evacuation at departure



(a) Predictor importance ranking

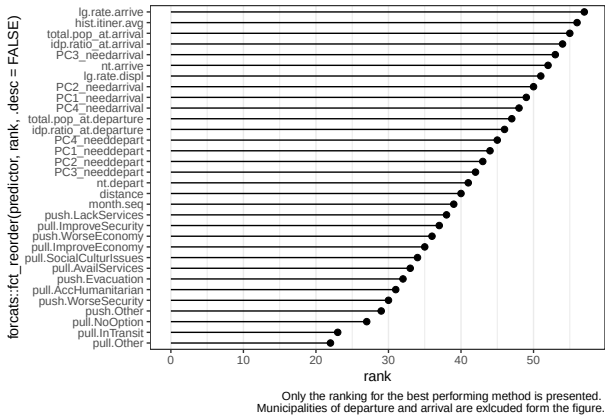
model	RMSE	Rsquared	MAE
LASSO	0.2012	0.9352	0.1566
PLS	0.2013	0.9353	0.1565
RF	0.0893	0.9880	0.0498

(b) Model evaluation parameters

Figure 14: Model 3 performance

- Push factor - Other, other non-specified factor at departure
- Pull factor - Improved economy, at arrival location
- Pull factor - Improved security, at arrival
- Pull factor - Availability of services, at arrival
- Pull factor - Access to humanitarian assistance, at arrival
- Pull factor - Social and cultural issues, at arrival
- Pull factor - In transit, IDP arrived in transit to other location
- Pull factor - No other option, there was no other possibility than to arrive at this location
- Pull factor - Other: other non-specified factor at arrival

The model accuracy is already outstanding and we could expect the addition of these factors would lay on very slight improvements on its performance, if any. However, these improvement prove that the addition of this factor improve predicting capacity fro IDP displacement. The RF was still and finally the best performing method. Regarding the importance of the new predictors, we observe that only one Pull factor - Improved economy position among the ten best performing predictors as average rank of the three models.



(a) Predictor importance ranking

model	RMSE	Rsquared	MAE
LASSO	0.2002	0.9360	0.1555
PLS	0.2002	0.9360	0.1553
RF	0.0907	0.9877	0.0509

(b) Model evaluation parameters

Figure 15: Model 4 performance

The best RMSE was of 0.0907 with RF method. Overall, most important predictors are still historical average displacement for the itinerary and Rate of arrivals (log)

4.6 The final test

Model 2 with RF method was the best performing for predicting IDP displacement. Meanwhile the inclusion of sector needs and reasons of displacement are likely to improve the baseline model, they do not provide gains over a model that comprises demographic predictors. We tested the model against our hold out data, which was fully excluded from data exploration and model training, and obtained an outstanding RMSE of 0.0607 in the logarithmic scale. Reverting the logarithmic transformation to its original scale, the RMSE was estimated in 1.15 IDPs.

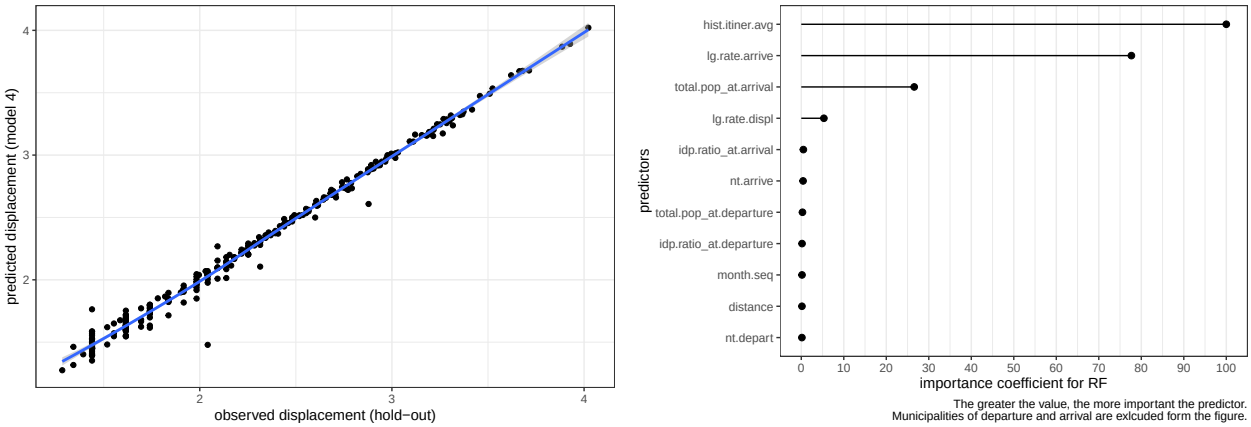
Table 4: Baseline model evaluation: Historical average of the itinerary

model	RMSE	Rsquared	MAE
Model 4 - RF	0.0607	0.9905	0.033

As we could see from Figure 16, the observed IDP displacement was tightly fitted by the prediction. Nonetheless there were few cases that significantly deviated form the regression line. This cases tend to be more on the lowest values of displacement, indicating our model would have greater difficulty on accurately predicting lower values than larger values. In future works, a look into the residuals could help to clean the model by removing them from the training and achieve even more accurate fitness.

When observing at predictor importance (Figure 16) could conclude with three main observations: 1) Three main predictors account for the greates part of the improvements in RMSE during RF iterations; 2) Predictors that capture the IDP displacement in its composition tend to perform better than those that are composed independently of IDP displacement; 2) Predictors at arrival tend to perform better than at departure, which could be explained due to the limitation is the collected data, vis-a-vis data on departure is not complete and only refer the most common location of departure for each location of arrival.

For a matter of computational efficiency, the whole model could be reduced tot he first four indicator and still preduce very similar RMSE values to the ones reported here. The complete list fo predictor importance can be found in Annex B.



(a) Regression fitness between observed and prediction

(b) Predictor importance for RF method

Figure 16: IDP rates and displacement

When drawing the time series of the prediction and the observed IDP displacement we can observe the two almost overlap; the model predict with high level of accuracy the trend (see Annex A for the same plot for the original scale of the outcome). Not surprisingly, the model is able to capture the spikes of displacement at provincial level over the training data series. What is more indicative is that it also capture the spikes for the two last month that were hold out from the training. Despite of this positive result, still will need to

wait more to observe if in the future the model is able to capture the spikes for new months to come. What we could be sure it would predict the stable base of IDP displacement.

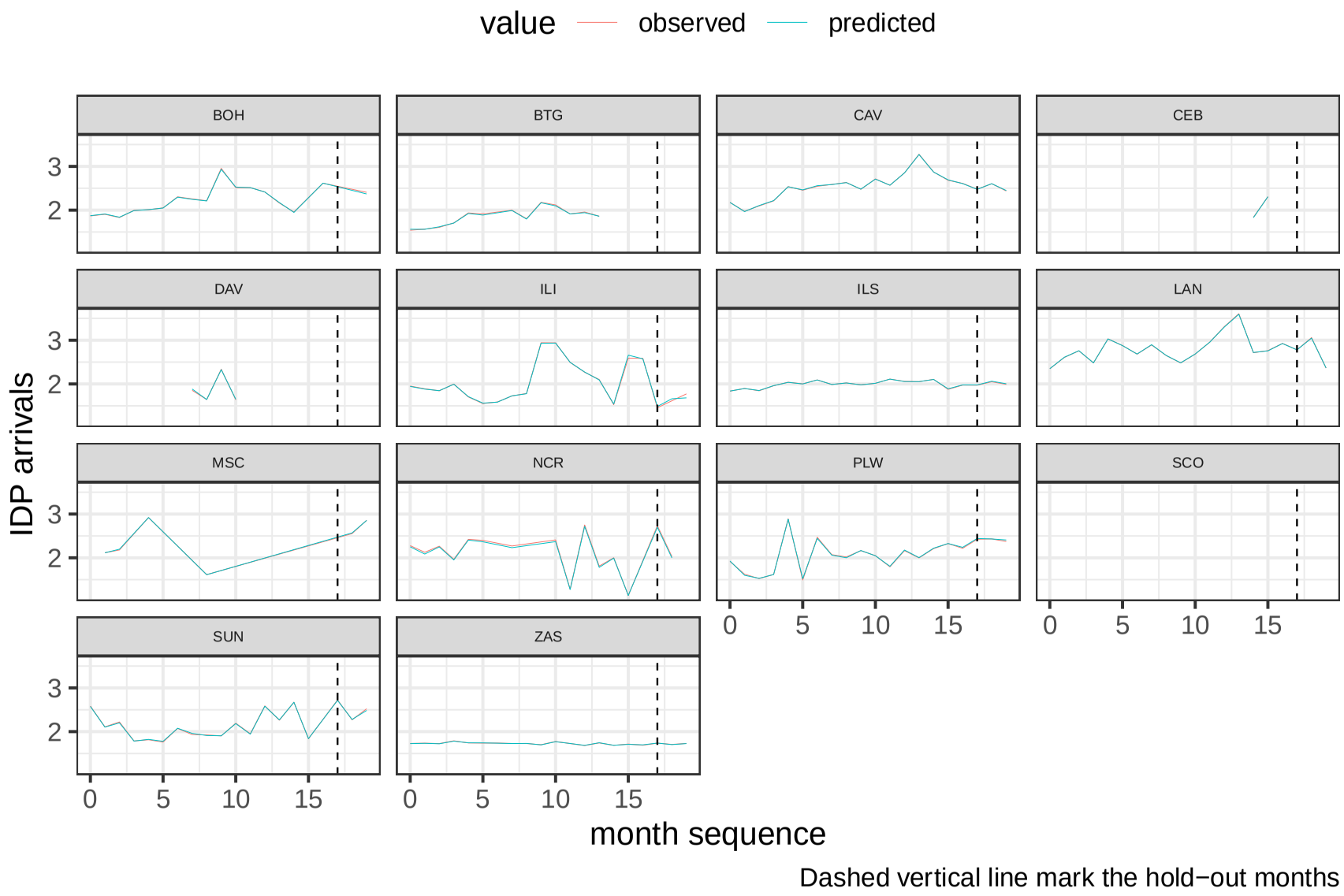


Figure 17: Observed and predicted trend of IDP displacement (log) by province of arrival

5 Conclusions

This study had allow us to develop an accurate prediction model for IDP displacement. Furthermore, we were able to see how meanwhile displacement variables are not enough to outperform the historical average, the inclusion of demographic predictors increases significantly the accuracy of the model. The inclusion of sector needs and reasons for displacement provide little gains, if any, into the model and can be left aside. The lack of improvement provided by this variables can be either inherent to the data collected or to the type of predictor derived from them. Other type of predictors to refer persons in need and pull and push factors might become more relevant than the ones here presented.

However, it could be indeed a desired side-effect. Data inherent to displacement and demography is the best perfoming in the model and this data is available for all itineraries, even those where there have never been displacement. This way, we could run the model in a complete data set of itineraries, including those where there have never been displacement (i.e. outcome = 0) to predict not only quantity of displacement by "if" there is displacement.

This prove that, for the purpose of predicting IDP displacement, there is no special need to load data collection with with additional variables on contextual factors. Although it will not harm the model, the efficiency of the gains would be questionable. However, it is indeed important to complement the displacement data with demographic variables.

We also believe that the availability of data at a low level such as the municipal level have allow the good performance of the model. Other studies, such as Anonymous (2018) have study the matter ar provincial level with significantly lowers RMSE (~0.65).

As more practical, outputs of the current study the prediction could be used to anticipate humanitarian response in the location of arrivals, at the same time that increase protection measure in the location of departure to prevent displacement and guarantee that the displacement happens with the greatest safety and dignity possible.

All of the above given, the model still have important limitations. We believe that due to the nature of the data we have input into the model, it is mainly capturing the stable base of displacement and will have difficulties on capturing massive IDP displacements stemmed from sudden critical events leading to crises. Meanwhile the model is still relevant in scenarios of protracted crisis, it might face this limitation on scenarios of active crisis.

At this point will be worth to explore the addition of other predictor related to information that capture this sudden and critical events. Although this information tend to be scarce and lagged on time, an appropriate manipulation might improve the weaknesses of the model for scenarios of active crises.

As additional set of predictor that would be relevant to test would be: market prices, conflict data and area of control in conflict scenarios, population compositions (gender and age). Furthermore, predictors that were not included due to computational limitations, but would be relevant to explore are: inclusion of itinerary it self and some compositions at itinerary level, such a number of time the itinerary as place of departure/arrival, the accumulated displacement in an itinerary, more accurate formulation of distance, such as instead of using municipality centroids, using average distance across all point to point itineraries in a municipality. Finally would be interesting to consider departure predictors with lag of one month, as although the data collected is not lagged, this might allow to count for the fact that departure occurs before arrival.

Another important limitation fo the model is that is only able to produce prediction for itineraries for which there have been already displacement. It cannot predict the displacement in new itineraries (i.e. itineraries that do not appear in the historical series). For this might be relevant to complement the current model with another classification model with a binary formulation to predict if displacement will occur on an itinerary or not. Then for all itineraries were displacement would be prediction, we could run the current model to predict the amount of displacement

Finally, across all the development of the model we have always give preference to data transformation that would improve the model's accuracy, rather than model's interpretability. Among two of the most

significant were the transformation of the outcome and some predictors in a logarithmic scale, and the matrix factorization applied to sector needs. Meanwhile logarithmic transformation are susceptible to be reversed back to natural numbers, matrix factorization cannot be reversed; however we tried to cover this limitation offering help for the interpretation of the components.

The present study was mainly focus on the technicalities of the modeling approach and less on the theoretical foundations of IDP displacement that it should guide it. Further literature revision and theoretical bases are needed to complement the study..

6 References

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. 2018. "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction." *Computational Statistics & Data Analysis* 120 (April): 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

UNHCR. 2018. "Emergency Handbook (4th Ed.)." *Emergency Handbook*. https://emergency.unhcr.org/.

7 Annex A: Observed and predicted trend in original IDP displacement scale

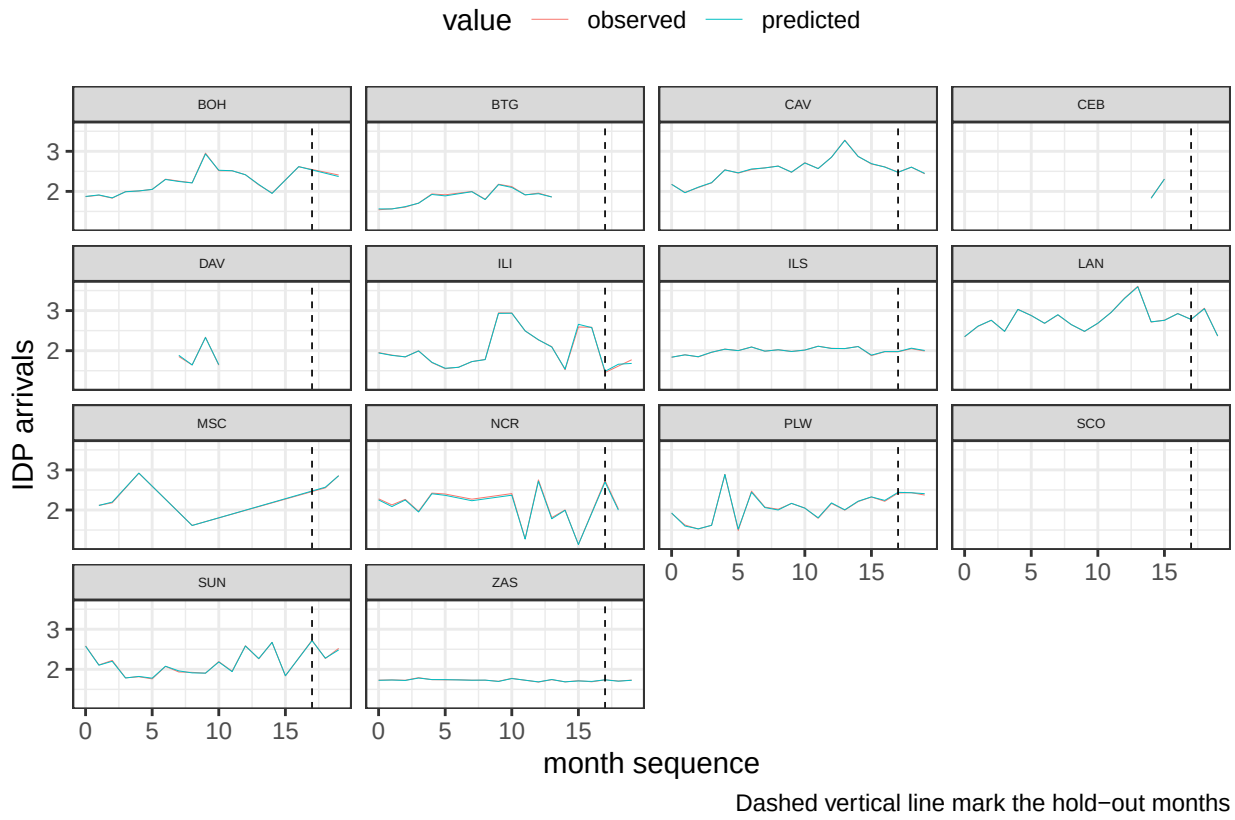


Figure 18: Observed and predicted trend of IDP displacement by province of arrival

8 Annex B: Predictor importance of the final selected model (Random Forest)

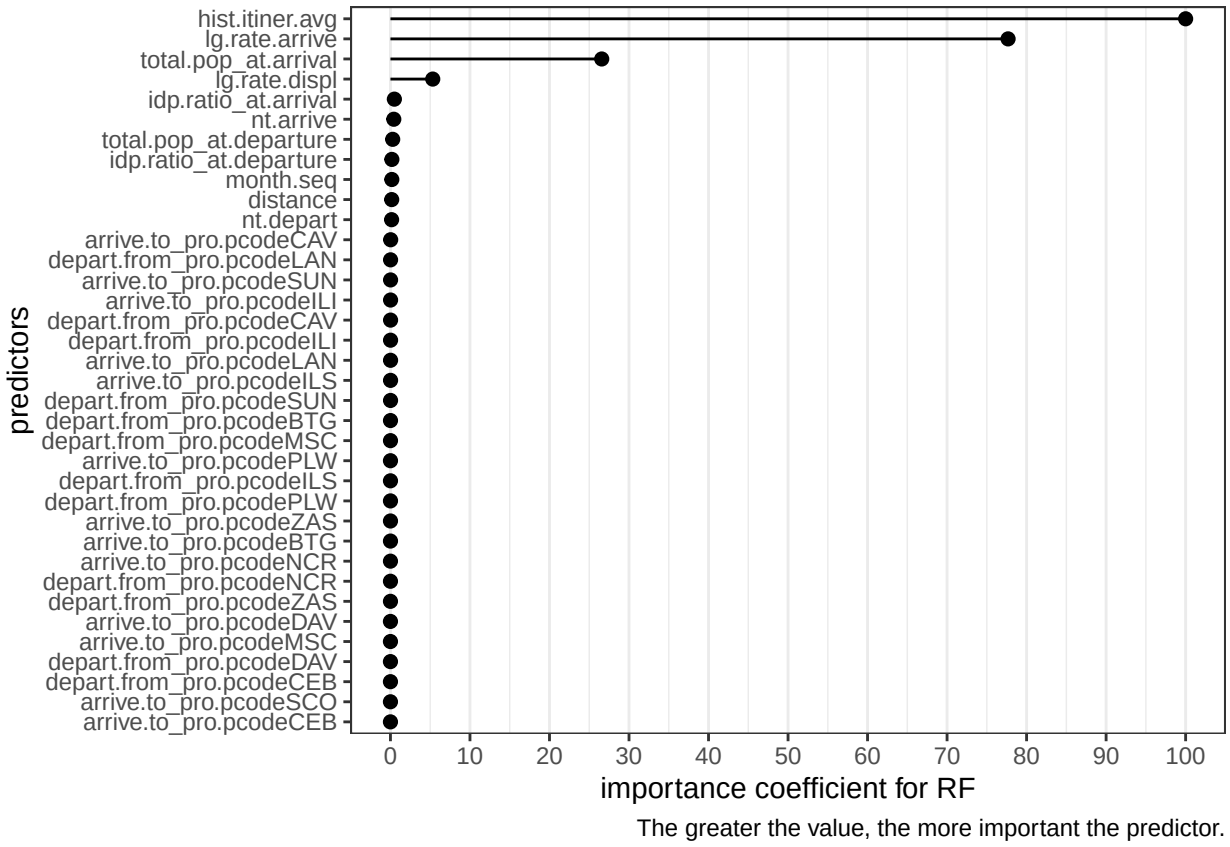


Figure 19: Complete list of predictor importance of the selected model