Real-Time Displacement Forecast for Natural Hazards

Janik Baumer, Vincent Bardenhagen, Daniel Benesch, Mian Zhong, Xiang Li, Gaël Perrin, Nathan Rouff, Anastasia Sycheva, Lionel Trebuchon.

Abstract—

This study presents a complete framework of real-time displacement prediction caused by storm and flood events. It allows to access and combine various data sources reflecting hazard intensity, exposed population, vulnerability and the people displaced in the past to train a machine learning algorithm that allows to forecast displaced people in future events.

Keywords—Natural Hazards, Displacement, Machine Learning, Geospatial analysis

1 Introduction

Natural disaster is one of the primary causes of *internal displacements*, forcing large part of the population to leave their homes or places of habitual residence. During the year 2018, it is estimated that 17.2 million of internal displacement were caused by natural disaster, where 5.4 and 9.3 million can be attributed to floods and storms, respectively [2].

In order to mitigate the consequences of internal displacement of such a large scale, it is crucial to predict the impact of a natural *hazard* (storm, flood) as soon as possible. This information is necessary to allocate local and international resources as quick and efficient as possible. In this documentation, a framework to estimate the number of people displaced for a given hazard is presented.

Section 3 reviews the dataset that was used to test and validate the framework. Section 2 gives an overview of the framework. In section 5, the performance of the framework is tested by comparing the predictions to the real values.

2 The Captain Framework

The developed framework includes all crucial steps of an analysis pipeline from data acquisition, matching of data sources, feature creation, machine learning analysis to prediction in a concise way. The data acquisition step downloads relevant hazard data, additional vulnerability characteristics and extracts the relevant information for further analysis. The matching of data sources includes matching in two aspects. Firstly, the displacement figures that are associated to a certain hazard event are found. This matching is based on multiple heuristics of timely and spatial proximity. Secondly, the exposed population is calculated by a geospatial matching between the area affected by a certain hazard intensity and the habitants of this region.

The feature creation sums exposed population over hazard intensities after binning and income levels and adds features of additional vulnerability characteristics. The machine learning step automatically analyses multiple models and parameter constellations and selects the best for the given dataset. The prediction is made based on the selected model.

3 Datasets

The analysis is based on four categories of data, *exposures*, *hazards*, *displacement* and *vulnerability proxies*, that are described below.

3.1 Exposures

The data for the *exposures* is from [?] and describes the population on a $5 \times 5\,\mathrm{km}^2$ grid disaggregated by building type in maximally 15 categories and income level in three categories (low, middle, high). The data reflects a population estimation from 2015. The data is available for Asia, Africa and Australia. For the Americas data was only available in the coastal areas on a $1 \times 1\ \text{km}^2$ grid.

3.2 Hazard

Storm and flood data originate from the "Global Disaster Alert and Coordination System" [3]. For each storm this data consists of multiple files including some meta information and coordinates of the outer borders of polygons. Each file describes the state of a storm at a certain point in time. For each file there are multiple polygons, and each covers the area where the windspeed was at least the value of the polygon label. In our analysis we aggregate the area covered by polygons over files and select the highest windspeed for each area.

For each storm this data consists of only one polygon with some meta information on the flood. The area covered by the polygons is an over approximation of the affected area including all affected regions. For our analysis we bin the flood intensity into four categories.

3.3 Displacement

For the displacement caused by storms we solely use [5] where the number of displaced people per country and event are given. Furthermore, there are some information on event name and event date, but the naming scheme and event data association given an event over a longer period is not consistent. For the displacement caused by floods we use [5] if available and if not available web scrape displacement information from the [3].

3.4 Vulnerability proxies

As socio-economic situation is closely related to resiliency to hazards were added some features reflecting this to the feature space of machine learning models. From the world bank [6] include the GDP of the country the hazard hits, its total population and surface area. From the [7] the socio-economic vulnerability and the evaluation of physical infrastructure were included.

4 Machine Learning

Our resulting data matrix gets split into a training (80%) and a test set (20%). We perform a randomized search cross validation with 10 cross validation folds and 10 sampled parameter configurations for XGBoost, Random Forest, Elastic Net and Ridge Regression. The best performing model was a random forest. Subsequently, we trained the best model on the whole training set and evaluated its performance on the held-out test set. This a reasonable generalization error estimate. To show our improvement over climada's prediction, we also evaluate the latter's prediction on the test set. We managed to reduce the root mean squared error from 10^6 to 10^5.

5 Results

The accuracy reached in the machine learning model is not enough to be used for efficient resource allocation yet. However, it already is a large improvement in comparison to the existing estimation techniques.

6 Conclusion

The forecasting of displacement by natural hazards is a challenging task as many factors need to be included. The main effort for this project was put in the data acquisition, data source matching and feature creation with great successes. However, the results are still in a review stage and the framework is not yet capable to run online and deliver real-time predictions in a user-friendly manner.

References

[1] Internal Displacement Monitoring Center, Global internal displacement database, http://www.internal-displacement.org/database/displacement-data.

[2] Internal Displacement Monitoring Center, Global Report on Internal Displacement, May 2019 http://www.internal-displacement.org/sites/default/files/publications/documents/2019-IDMC-GRID.pdf.

[3] Global Disaster Alert and Coordination System, https://www.gdacs.org/Alerts/default.aspx.

[4] ISDR (2015) Global Assessment Report on Disaster Risk Reduction, United Nations, Geneva, Switzerland, http://unisdr.envcomp.eu/

[5] Internal Displacement Monitoring Centre, http://www.internal-displacement.org/database/displacement-data.

[6] The World Bank, https://data.worldbank.org/indicator/

[7] InfoRM by the Inter-Agency Standing Committee and the European Commission, https://drmkc.jrc.ec.europa.eu/inform-index/InDepth/Publications