Hack4Good: Report Team Green

Stephan Artmann, Viktoria de La Rochefoucauld, Nico Messikommer, Francesco Saltarelli

June 7, 2019

Abstract: *An important objective of humanitarian aid is to identify households in need. This is usually done by reaching out to people to answer a questionnaire detailing their living situation. Based on their answers, sectoral Index Indicators (PiN) are calculated. This is a time consuming and labor intensive process. We present here a different approach aiming to predict the PiN on demographic variables. This approach drastically simplifies the process. Additionally, as we show, it can be significantly improved with other publicly available development indicator datasets, e.g. population density and poverty. Furthermore, we provide a prototype of a visualization app to further help identifying people in need. To improve both approaches, we recommend to document each households‘ geographical location in a more precise manner in further studies and to provide a formula to calculate the PiN-score.*

1. Introduction

In 2018, IMPACT reached out to households in Nigeria to answer a questionnaire detailing their living situation and needs in different aspects of life. In this report, we outline our efforts to analyze the questionnaire data, which were two-fold. First, to help identify households in need based on (easily obtainable) demographic data rather than expensive and time-consuming questionnaires. Second, to provide a framework for the interactive visualization of the given data on a map.

This report is structured as follows: Section 2 gives an overview over the questionnaire data provided, some limitations we encountered and how they might be resolved in the future. In Section 3, we detail our method used to calculate the PiN. Sections 4 and 5 are dedicated to PiN-prediction and the visualization app, respectively. Finally, in Section 6, we give an outlook on how to improve predictions, expand the visualization app and detail what skills might be needed to continue working in this direction.

2. Questionnaire Data

Questionnaire results were provided by IMPACT in form of two tables, one depicting answers regarding a household (‘household-table’) and another containing data regarding the household members (‘hh-member’-table). Furthermore, there was a table including location data for each settlement camp (‘initial’ table). As a first step towards our analyses, we generated a ‘clean’ data table containing demographic variables, the first three priorities each household stated, and the camp the household was located in (as well as the Camp‘s coordinates). We stress, though, that matching households to their approximate location from the tables provided was difficult and may probably be improved in the future. For a detailed description of the clean data table and how it was generated, see the README.md file in the gitlab-repository.

3. PiN Calculation

We tried to follow the guideline provided by IMPACT to calculate the PiN-values, see the REAdME.md-file in our repository for details. This way of calculating PiN-scores seems to be rather error-prone, though. When comparing the calculated PiNs of the four participating teams, we found that they all differed noticeably. This may be due to mistakes in calculation, but also due to inconsistencies in the data, as the following example illustrates: The PiN-score of a household depends on whether or not there are barriers hindering access to food from a market. Surprisingly, there were households which apparently answered ‘Yes’ when asked whether no such barriers existed, but when asked for specific barriers (market too far away, food too expensive etc.), claimed that some of them did exist. The PiN-score was therefore not well-defined for such households. As a consequence, we strongly suggest to provide PiN-scores or a unique way of calculating them in the future.

4. Prediction of Households in Need

The first goal of our project was to classify, if a Household is ‘in need’ in a humanitarian sector based solely on the demographic variables. To tackle this task, we decided to evaluate the performance of a logistic regression model as well as a random forest model. The choice of both models was motivated by the fact that both are commonly used in the literature and were successfully applied for similar tasks.

In the logistic regression model we employed a L2 regularizer. The strength of the regularizer was tuned for each sector separately by minimizing the F1 score of the predictions of the class ‘not in need‘. In contrast, the random forest model was tuned based on the overall performance on all sectors. The number of trees was fixed to 50 with no depth limitation. As the datasets were unbalanced, the class weights were computed based on the number of samples in the specific class.

The extracted dataset was split in a training and a test set. The parameter tuning of both models was performed on the training dataset by using 5-fold cross-validation. The performance of both models on the test dataset for the task of predicting whether a household is in need or not based on the demographic variables can be seen in the Table 1. The nutrition sector was excluded from our experiments as the two classes in this sector were highly unbalanced.

In a subsequent step, we wanted to show the influence of additional data sources for the classification task. As a first idea, we wondered whether there might be a correlation between the ‘in need’ classification of a household and the poverty and educational background of a household as well as the population density at the location of the household. Hence, we decided to incorporate high resolution maps of Nigeria containing these three development indicators provided by the WorldPop project (https://www.worldpop.org/). The maps are illustrated in Figure 1. For each household, the three additional indicators were extracted based on the GPS coordinates of the household. The three development-indicator datasets as well as other spatial demographic datasets can be found on the WorldPop project website for a large number of countries around the globe. Note that as the additional data was extracted based on the location of a household, we assumed a spatial correlation between households as well.

The additional input improved the F1-Score for almost each sector for both models. Especially the performance of the random forest was increased by a significant margin. These results show that it is worth considering adding easy accessible data to the input for the prediction task. Moreover, a finer location information of a household could lead to more accurate indicators extracted from the high resolution maps resulting in a better classification performance.

Model	Wash	Shelter	Food	Livelihood	Health	Education	Protection
Logistic Regression	**0.65**	**0.62**	0.65	0.59	**0.67**	**0.63**	0.67
Logistic Regression+	0.64	**0.62**	**0.66**	**0.60**	**0.67**	**0.63**	**0.72**
Random Forest	0.85	0.51	0.43	0.58	0.62	0.67	0.9
Random Forest+	**0.9**	**0.6**	**0.53**	**0.66**	**0.69**	**0.73**	**0.93**

TABLE 1: Performance of the classification models. F1-Score for each sector corresponding to the ‘not in need‘ class. The ‘+’ sign after the model name indicates that the development indicators were employed as additional input data.

5. Interactive Visualisations

Data analyses and predictions can be challenging to interpret, let alone to obtain an intuition for the described phenomena or prediction results. Therefore, we aimed at providing IMPACT with a versatile visualization framework, which can be used to showcase the status quo and compare it with the models‘ predictions. This helps in identifying regional patterns.

We developed our code in Python and based the visualization app on the library Bokeh. This library empowers the Python user to generate visually appealing plots and standalone web-applications by handling all necessary Javascript and CSS in the background, requiring little to no Javascript knowledge from the latter. We additionally employed Pandas-Bokeh, which is a higher-level library written to further simplify the usage of Bokeh, by hiding necessary pandas-to-bokeh data source conversions from the programmer.

In order to create the basis for the interactive map, we used the publicly available OpenStreetMap data, which is integrated into Bokeh as a shape source for the geographical plots. Then, we gathered geographical data outlining the wards in northeastern Nigeria and averaged the calculated and predicted PiN results with respect to the wards, as well as the individual villages that lie within the wards. When combining these pieces of information into map format, it was crucial to pay attention to latitude/longitude to Mercator coordinate conversions, including possible offsets. We found the Pandas-Bokeh implementation to be the easiest to reproduce and have therefore also provided this solution in our codebase. However, in order to have full control over the interactivity, which in our case entailed the hover tool and dropdown menu, it was necessary to alter a custom Javascript block, which is located inside the Pandas-Bokeh library and did require some Javascript knowledge.

Finally, it is possible to manually or automatically launch an interactive version of the plot and interact with it, both within the jupyter notebook environment as well as in a web-browser-based format.

6. Outlook and Recommendations

We believe that performance of both approaches could be increased by improving documentation regarding which column in the data tables correspond to each other (in particular to help calculate the location of each household), and by providing an exact formula for PiN-scores.

We encourage future users to see the map we provided as a starting point towards building more complex visualisations. With Bokeh, it is possible to link plots together, build dashboards and even whole storyboards to convince a fellow colleague or funding body of the needs in a specific sector. For instance, starting with the map we created, we could empower the user to tap on a specific ward and thereby trigger the update of a series of other plots next to the map, which show e.g. the distribution of PiN scores. These plots can then be broken down by population group of a different demographic factor, population density and other important variables. Further, incorporating additional data sources can enhance not only the predictions, but also the visualizations, providing the user with an increasingly powerful tool. Finally, the entirety of plots and maps can be launched in the form of a standalone web-application, which is distributable among peers. This can empower the latter to gain a more thorough understanding of the data by interacting with the plots him/herself and be part of a convincing argument for IMPACT.

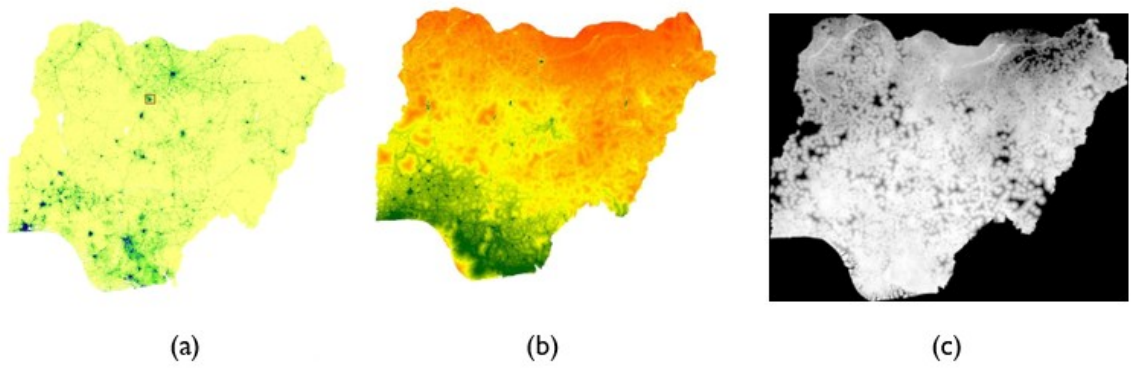

FIGURE 1: (a) Estimated population density for the year 2020. (b) Proportion of residents living on $1.25 a day. (c) Proportion of male literacy.