Hack4Good Pilot Edition: Assisting Data-Driven Decisions in the Humanitarian Crisis in Nigeria

Marco Mancini, Yilmazcan Ozyurt, Ylli Muhadri and Maria R. Cervera

***Abstract*— In order to assist decision-making in the humanitarian crisis in Nigeria, we tried to identify undiscovered patterns on the Multi-Sectorial Needs Assessment dataset collected by REACH. First, we developed a random forests model that was able, to a low degree, to predict the overall level of need of a household. Second, we identified sets of co-occurring sectorial needs. Third, we showed that the current methodology doesn't allow to accurately predict the reported needs of the households. Overall, these results shed light on the nature of the collected data and suggest directions to improve the assessment methodology.**

I. INTRODUCTION

A decade ago, an armed conflict led to a humanitarian crisis in the North-East of Nigeria, which lasts until today. The crisis has caused the displacement of millions of civilians, and it is estimated that seven million people are in need of humanitarian assistance. In this context, REACH conducted in 2017 a Multi-sectorial Needs Assessment identifying the needs for internally displaced peoples, returnees and non-displaced peoples. This dataset, which combines information on a variety of sectors including Food, Education, Health etc. offers unique potential to investigate underlying relationships between the characteristics of the people and their needs, ultimately helping decision-making for humanitarian aid. Here, we tried to further explore this dataset and uncover patterns so far not identified.

II. METHODS

Datasets were provided by REACH. Analyses were done with *Python 3.5* using *numpy, pandas, scikit-learn, yellowbrick*.

A. Datasets

The data collected in a survey on households (HH) in accessible areas affected by the conflict. It was provided as two main clean *csv* files, each containing both demographic information and answers about the situation and needs of the HH. The first consists of HH-related information, and the second consists of individual HH members.

B. Data Pre-Processing

We created two matrices with a reduced number of HH demographic and need features, respectively. For both matrices, we only kept HHs that consented to the interview (10 378). For the demographics dataset, only features relating to geographical location, number and age of HH members, languages, and respondent and household head information were used for the analysis. Categorical features were transformed into dummy variables, information about languages was combined into a single numerical feature specifying the number of languages spoken, and binary features were converted to booleans. The final number of features was 35. For the sectorial needs dataset, we combined the answers to the survey relating to the particular situation and needs of the HHs into a set of sectoral severity scales. The severity score for each of the eight sectors was calculated following the sectoral composite indicators datasheet provided by REACH. There were 8 features (sectoral severity scale), each with need values from 0 to 10.

C. Machine Learning Methods for Studying the Relationship Between Demographics and Needs

Our first goal was to identify relationships between the needs of the HHs and their demographics. First we computed an index of needs vector, representing the overall level of need of the HHs (*no needs, low needs, medium needs, high needs or very high needs*) according to REACH's definition. We then tried two approaches: predictive models and clustering. Several predictive models were tested, from simple linear classifiers to complex ensemble and boosting methods. We focused on Balanced Random Forest (BRF), which if well suited for tasks with imbalanced classes. For clustering the HHs according to their demographics, we used Gower dissimilarity measure [1] to introduce the dissimilarity matrix of HHs based on the features, and then applied Constant Shift Embedding (CSE) [2] to project our matrix into 2D space for visualization. This approach allowed us to combine both numeric and categorical features.

D. Sector Clustering

To exploit the unique potential of the available multi-sectorial data, we investigated whether sectors could be clustered together. The goal was to determine whether some HHs experience needs in given sets of sectors, rather than individual sectors. We performed feature agglomeration on severity scores after normalization. By clustering with a progressively smaller number of clusters, we sketched a tree showing the relationships between the clusters.

E. Critical Evaluation of Need Predictions

Last we investigated whether the indicators used to calculate the severity scale for each sector are optimal. Specifically, we compared the sector with the highest severity score with the reported first priority of the HHs. We used confusion matrices for evaluation. We also used principal component analysis (PCA) and CSE to verify that the obtained clusters were consistent across methods.

III. RESULTS

A. Relationship Between Household Needs and Demographics

The index of needs' classes are highly imbalanced. The number of HHs per class was: no needs 253, ow needs 2375, moderate needs 5032, high needs 2617, and very high needs 101. Besides, mutual information and correlation scores between the needs and the demographics were low, suggesting

no strong dependencies to learn from the data. Of the predictive models tested, BRF worked better than the alternatives, with a balanced accuracy of 41%. The results of the CSE clustering can be found in Figure 1. We cam see that sectorial needs

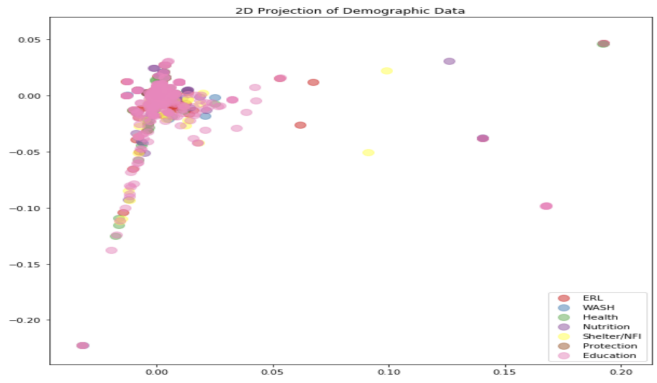

Fig. 1. Demographics data projected in 2D after Constant Shift Embedding

overlap, showing no distinct HH clusters with different types of needs.

B. Sector Clustering

Figure 2 shows the tree structure obtained by clustering the sectors. We see that needs in the Food sector often co-occur with needs in the Early Recovery and Livelihoods sector. Furthermore, these are more often related to needs in either Shelter or Protection, than with other sectors. Wash and Health also often co-occur. Notably, PCA and CSE lead to similar results, indicating the robustness of the results.

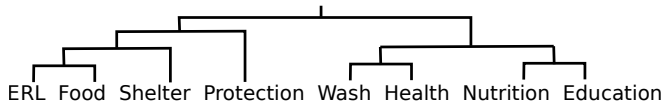

Fig. 2. Clustering of the sectors

C. A Closer Look at the Methodological Assumptions in Multi-Sectorial Analyses

Figure 3 shows the results of the comparison between the sector with the highest severity scale and the first need reported by the HH. We see that the first priority of the HH is rarely

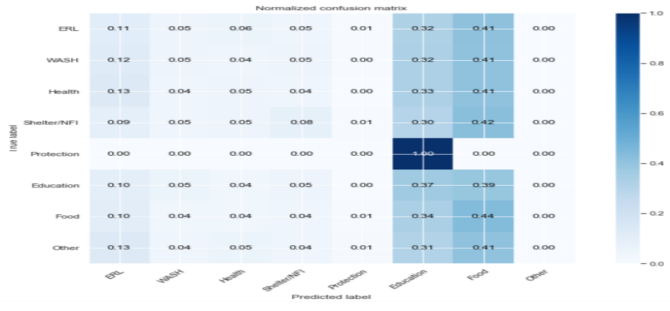

Fig. 3. Normalized confusion matrix of the HH's first need prediction using the provided severity scale indicators, compared to the HH's reported first need. Numbers show the true positive rate.

correctly identified. Particularly, all the HHs that reported Protection as their first priority were labeled as in need of Education. Generally, priorities in Education and Food are over-predicted.

IV. Discussion

In this project, we tried to identify possible relationships between the HHs needs and demographics. Using BRFs, we obtained above-chance classification of the index of needs of HH, although the performance was low. Clustering the HHs according to their demographic data did not produce clusters with characteristic needs patterns. These results underline the challenge in trying to identify an underlying data structure, but don't rule out that different approaches, for example trying to predict other measures than the index of needs, could lead to improved results.

The sectorial clustering showed close relationships between several sectors, such as Early Recovery and Livelihoods with Food, Health with Wash. The results can be intuitively understood: a HH that has economic problems (ERL need) is likely to have problems to provide enough food, just as a HH that has limited access to clean water and sanitation will be more likely to develop health problems. Our results differ significantly from the common combinations of sectors in need reported in REACH's report [3], where Education-Health, Food-Education and Food-Health were the most commonly occurring combinations. This is however easily explained by the fact that those sectors are the ones most frequently in need in absolute numbers.

The critical analysis of the sectorial severity scale calculations showed that in the vast majority of the cases, the first need reported by the HH is not correctly predicted by the given indicators. Although it could be argued that the HHs might be biased in some of their answers, we believe that such a strong deviation suggests revising the methodology used to compute the severity scale.

V. Conclusion

Using multi-sectorial data from households in affected regions of Nigeria, we tried to uncover patterns in the data that could assist decision-making to tackle the humanitarian crisis. Our results stress the difficulty in trying to establish clear relationships between the demographics of the households and its level of needs, but don't exclude the possibility that such relationships can be established. Besides, we found that households were likely to have needs in given sets of sectors simultaneously, which could be of importance to design mitigation strategies providing combined help in given sets of sectors. Finally, we showed that the current severity scale metric doesn't reflect the reported household priorities, indicating that the methodology might need to be revisited. Overall, we expect these findings will be of use for exploiting with multi-sectorial data in the humanitarian sector.

References

[1] Gower, John C. "Some distance properties of latent root and vector methods used in multivariate analysis." Biometrika 53.3-4 (1966): 325-338.

[2] Roth, Volker, et al. "Optimal cluster preserving embedding of nonmetric proximity data." IEEE Transactions on Pattern Analysis Machine Intelligence 12 (2003): 1540-1551.

[3] REACH, Nigeria: Multi-Sector Needs Assessment. March 2019.