Bayesian Networks for Speeding Up Data Collections

presented at Hack4Good on May 20, 2019

Martin Buttenschön Natallie Baikevich Luca Pedrelli Georgios Papadimitriou

***Abstract*—Data collection and analysis is of great importance for humanitarian aid decisions. This is a challenging problem because obtaining a general yet clear picture of humanitarian needs in crises regions is labor intensive and costly. We explore the idea of using a generative model to decrease survey time. We show that using a Bayesian Network, one could pick questions during the interview to maximize information gain under a . This approach could potentially save time for first hand data collection or alternatively allow to collect more data points.**

I. Introduction

Interviewing people in extreme situations like an armed conflict or a natural disaster is a difficult and potentially dangerous task, especially given that the available time to collect the information is likely to be limited and, at the same time, NGOs need to act fast in order to have an effective impact on the situation and bring real help to the people, and what is even more important - the kind of help that they need. We propose a model that helps the user to prepare and formulate the right questions when interviewing people to understand their needs better.

II. Approach and Assumptions

We investigate the data collected in order to analyze humanitarian needs in Nigeria. To simplify the analysis we remove from the dataset the cases where respondents gave no consent to the interview, as well as some of the columns (e.g. time). In addition to that we reverse numerical encoding of categorical features (e.g. "yndk" questions or one-hot encoding). To facilitate the process of data exploration, the feature set is extended with sector indicators, cluster labels for each sector, and dependency ratios.

A. Multi-Sector Needs Assessment

Multi-Sector/Cluster Initial Rapid Assessment (MIRA) Analytical Framework suggests to evaluate the severity of humanitarian needs of households in multiple independent sectors: Wash, Shelter, Food Security and Livelihoods (Agriculture), Early Recovery and Livelihoods, Health, Nutrition, Education, and Protection.

The interview answers are summarized for each of the sector into *severity scale* (also called *weight*) ranging from 0, being the least severe unmet need, to 10, being the most severe unmet need. Moreover, a household scoring 4 or greater on the sector severity scale is considered to be in need of sectoral support. The number of sectors in need is often used to rank the households on the scale from "no needs" (needs in 0 sectors) to "very high needs" (7-8 sectors).

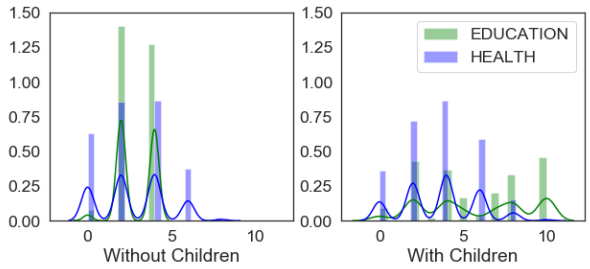

Figure 1. Need analysis in education (green) and health (blue) for families with and without children.

We incorporate the sector indicator features in our analysis. Note, that the provided data is not always consistent and in some cases ambiguous.

B. Data Exploration

Our analysis shows that there are general trends between demographics and people's needs. For instance, people living in similar family structures, i.e. presence of children or elderly persons in the household, face similar challenges. This is shown in Figure 1. Here, the scale goes from 0 (no need) to 10 (extreme need) for the sectors education and health. Note that health needs have are not significantly different for households with or without children. However, education need scores differ greatly between households with and without children. Based on these plots, we postulate a connection between demographics and needs.

Further, we explore the spatial correlation looking at the people needs on the ward level. Figure 2 shows the geographical area covered by our data in case of wash need score. We can distinguish a larger yellow cluster in the south part of the map where the wash need score is high. We use the coordinates of the centers of the wards to perform Moran's I test for the global spatial autocorrelations and confirm that for majority of sector indicators it is significant. However, for this kind of analysis, further investigation is required, since there are missing data due to unreachable wards and mismatches between the ward label in the interview data an administrative ward name used for the geographical map that leads to missing results on the plotted map[1]. We also

[1]The source of geographical data used for visualization is https://data.humdata.org/dataset/nga-administrative-boundaries

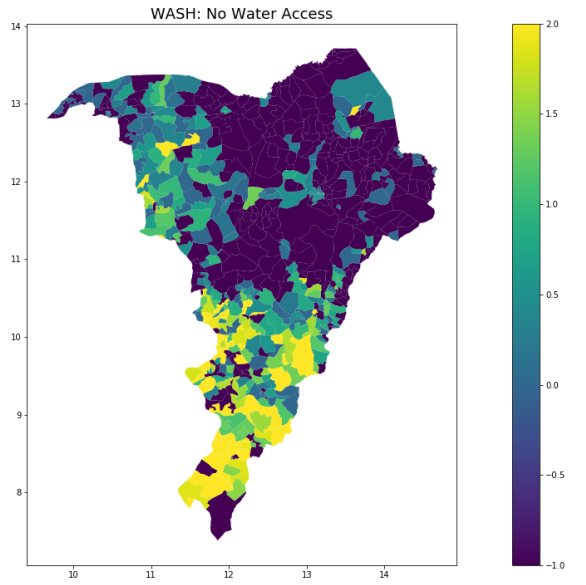

Figure 2. Spatial correlation in need in the wash sector.

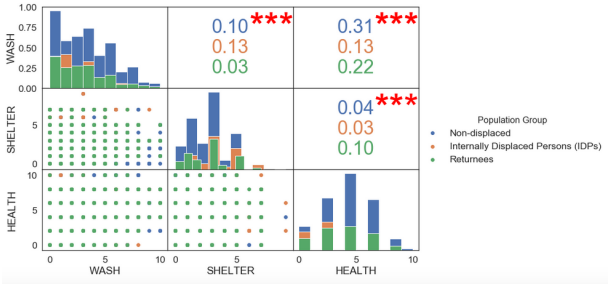

Figure 3. Example of inter-sector correlations.

recommend to perform spatial analysis on a finer level, e.g. settlement camps, but the data available doesn't allow to clearly determine the geographical location.

Finally, we challenge the assumptions of the MIRA framework (Section II-A) based on the fact that the sector needs are not independent. For example, the correlation between the wash and health sectors reaches 0.31 for non-displaced population, this value is not very high but it is statistically significant as shown in Figure 3.

C. Visualization

Data visualization was actively used in the exploration phase of the project by means of Python Notebooks and Tableau. The aggregated data is available in .csv and .pkl formats and can be mapped using LGA- and Ward-level shape files via LGA_Map and Ward_Map columns respectively.

III. Model

The similarities between households, their location and needs can be leveraged to develop a model, which helps the user to follow the right path when asking the questions by maximizing the acquired information under time constraints and avoiding redundant questions.

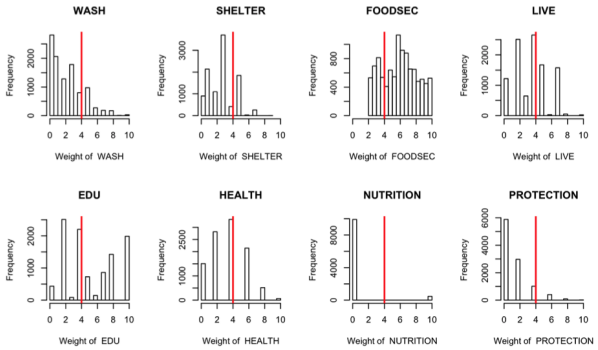

Figure 4. Distribution of sector needs indicators.

A. Model Structure

The model is structured as follows.

1) Data Preparation: Based on initial data preparation discussed in Section II, we transform each of the 8 sector indicators (on the scale from 0 to 10) into binary variables using 4 as a cutoff value. Those variables represent whether there is a need (above 4) or not (below 4).

2) Building Bayesian Network: The next step is learning the graph structure for Bayesian network with max-min hill climbing followed by a parameter estimation using maximum likelihood. We use the library 'bnlearn' in R for training [1].

3) Prediction: Finally, we can determine the right question to ask. When there is no observation available for a particular household, the best guess whether or not it has a need in a specific sector is the marginal probability (see Figure 4) of each variable. However, as information becomes available through the interview, we can update these probabilities in the Bayesian network and use these to choose what question to ask next in order to maximize the information gain (entropy).

B. Using the model

Let us illustrate the application of the model with an Example. If the interview is with someone with a wash need, the model predicts a high probability of health need. In this case, we could guess that they have a wash need and ask other questions first. On the other side, if a person has a no wash need then there is a 50% chance that the person has a need in health. This is the most uncertain we can be, and we should ask questions regarding the wash need.

This toy model shows in principle how to choose the next question category aimed at identifying household sector needs (Figure 5). A full model should ideally be based on the full battery of questions from the questionnaire.

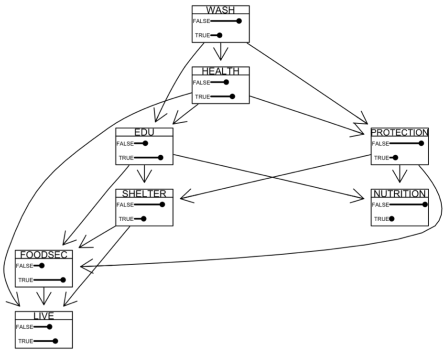

Figure 5. Bayesian network with conditional probability tables for selecting the next question category.

C. Further Work

Given the short timeframe and scope of the assignment, we showed that Bayesian Networks could be used in order to make the crucial step of question surveying more effective and less time consuming. Further work is needed in order to better understand the relationships between the various variables, the limits of the model, and also study datasets from other humanitarian initiatives.

Our model represents a trade-off between detailed data collection and time constraints/costs, thus, it is necessary to define what level of uncertainty is acceptable. For instance, it might be clear that household has need in health sector, but a detailed disease history profile is required for another research. Moreover, switching from sector to sector just because it is optimal from informational gain point of view can be confusing for an interviewee if the questions don't form a logical sequence.

Another aspect is technical feasibility, such as estimating the possible size of the graph and required number of interview responses for training a reliable model. While the training based on the sector indicators is quite straightforward, building a stable representation for the entire questionnaire is a much more challenging task.

Lastly, there has to be continuous input and guidance from the actual field workers, in order to eventually design a model that is usable and offers added value to the people on the field.

IV. Conclusion

We proposed Bayesian Networks as a statistical model to improve and accelerate data collection in humanitarian initiatives. Further steps are needed in order to confirm the feasibility and application of this proposal in the actual field. We show that in principle, one could use a statistical model to pick questions during an interview to maximize information gain under a time constraint. Despite the early stage of the idea, we are confident that this approach can save time for first hand data collection or alternatively allow to collect more data points.

References

[1] M. Scutari, *Bayesian Networks : with Examples in R*. New York: CRC Press Taylor & Francis Group, 2014.

[2] R. Nagarajan, *Bayesian Networks in R : with Applications in Systems Biology*. New York: Springer, 2013.