Discovering underlying Needs - Going beyond traditional Sector Analysis

Shirzart Enwer
ETH Zürich
aniwax@student.ethz.ch

Belinda Müller
ETH Zürich
bmueller@student.ethz.ch

Swaneet Sahoo
ETH Zürich
swaneets@gmail.com

Abstract

The traditional approach of Multi-Sector Need Assessment structures questionnaires for assessing people in humanitarian crises into predefined sectors. As a consequence, these questionnaires ... redundancy due to a substantial overlap between the sectors in terms of the needs of the interviewed households. We found that about 80% of the information captured by a set of preselected questions can be recovered from only four latent factors. We showed that these factors show a high linear dependence with some of the sectors, stressing the importance of these particular sectors. The exact semantics of these latent variables, however, may go well beyond the traditional sectors and is an open topic for future research. Thus, we suggest that re-thinking the traditional sector approach might lead to a more concise data acquisition and analysis .

1 Introduction

According to the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), the traditional tool for assessing the need of people affected by crises is Multi-Sector Need Assessment (MSNA). Information about the affected households is collected by means of interviews and then clustered into 8 sectors, as displayed in Figure 1[1].

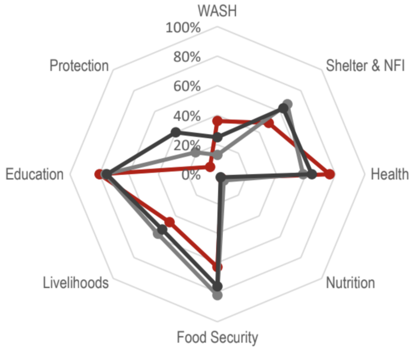

Figure 1: Multi-Sector Need Assessemnt in Northern Nigeria. The data collected from every household are classified into the 8 sectors shown above.

As part of the MSNA, the data is used to calculate person-in-need (PiN) scores for each sector, which indicate how strong the need of a specific household in that sector is. While the MSNA provides a comprehensive and detailed overview about the household-in-need situation in the region, we suggest that there is one major drawback associated with this traditional sector analysis: the questions offered in the interviews are not independent of each other, as often multiple questions point to the different aspects of a certain underlying need. Especially, questions asked for different sectors might correspond to the same underlying need. This introduces redundancy into the data, which makes both the interview and data analysis unnecessarily laborious. To discover the real underlying need, we suggest that the questionnaire for the interviews be re-designed such that the information captured by the questions is uncorrelated and redundancy is thus eliminated.

2 Methods

We used the sectorindicator.py from Team Black / PISA to calculate the PiN scores. It selected 34 features and created linear combinations, the PiN scores, for each sector except for nutrition, and an

[1] retrieved from http://www.reachresourcecentre.info/system/files/resource-documents/reach_nga_msna-brief_feb2019.pdf on June 7, 2019

overall PiN score as the sum thereof. We converted the CSV file to a binary matrix where the cells are binary, indicating whether that particular household's answer was indicative of the PiN score. Our task was then to find a simple mapping that compresses much of the binary matrix to a few new but meaningful factors. To obtain the latent factors of the data matrix, we used a variational autoencoder. We opted for this approach because it is a neural-nework-based compression method, which means, it allows us to exploit the non-linear power of neural networks, as opposed to principal component analysis for example. We tuned the network architecture and the number of latent dimensions by assessing the reconstruction error and found that the network with 1 hidden layer with 30 neurons and 4 latent dimensions was the most parsimonious one. Aiming for interpretation of the latent distribution, we correlated the four means of the latent marginal distributions with each of the 8 PIN scores, assuming that they will most likely yield the semantics for the latent space. We split the data into 2/3 training data and 1/3 validation data.

3 Results

Our model yielded a reconstruction error of about 18% on the validation set, that is it accounts for about 82% of the information encoded in the data set of 34 features with only 4 latent variables. Comparing this approach to a PCA, the latter only reproduces 62% with 15 factors already. The correlation of the means of the latent marginal distributions with the PIN scores is displayed in the correlation matrix in Figure 2. We observe that the first latent variable has a high correlation with education (0.78), the second with shelter (0.49) and protection (0.52), and the third with health (0.71) and WASH (0.54). The forth variable shows low to moderate correlation with most of the PIN scores but cannot be associated with a subset of these.

4 Discussion

We have discovered that about 82% of the information captured in the selected questions of the MSNA in Nigeria can be compressed to 4 latent factors which are strongly indicative of the underlying needs. We found high correlations of these factors with some of the traditional MSNA sectors, providing a first idea of the semantics of these factors.

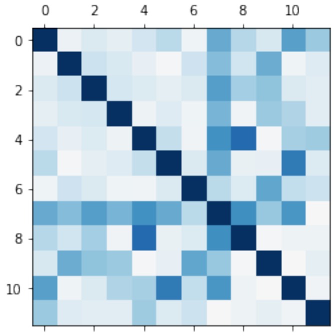

Figure 2: Matrix indicating the absolute correlation coefficient colour coded as ranging from 0 (white) to 1 (dark blue). The first 8 rows and columns correspond to the PIN scores as follows: WASH (0), shelter (1), food security (2), livelihood (3), education (4), health (5), protection (6), and overall PIN score (7). The remaining 4 rows and columns correspond to the means of the latent marginals.

This low-dimensional representation is a crucial step in improving the current questionnaires towards concision. However, this improvement can only be implemented when we know how to interpret the found factors and then can ask the corresponding questions. The found factors are nonlinear combination of the original questions, concretely, each latent factor can be traced back to an approximate second-degree polynomial of the questions. While it is technically feasible, it is conceptually non-trivial to identify the new questions, that generate the same answers as these combinations. A deeper analysis of the variational autoendcoder will likely reveal the few important questions for each latent factor. Here it will be imperative to not only consider correlation but also non-linear dependencies with potentially related variables. Also, it might prove beneficial to include other features, for instance demographic information, into the analysis. Lastly, since variational autoencoders are generative models, investigating the modeled latent distribution through the generation of representative data might yield further insights and thus merits future research. Our results might be complemented by an adaptive interview approach, where the question asked depend on the answer to the previous question.

Acknowledgements

We greatly thank Martin Barner, Chelsea Chen, Dr. Barbara La Cara, Simon Mathis, Sharon Orengo and Laetitia Vionnet for their fantastic support. Also, many thanks to ETH for funding this program.