EVALUATING NORWEGIAN CHURCH AID INTERVENTIONS THROUGH MULTIVARIATE ANALYSIS WITH BIG DATA

EMERGING FINDINGS ON HOW TO PAIR DATA FROM CIVIL SOCIETY ORGANISATIONS WITH EXTERNAL DATA FOR BETTER EVALUATIONS

Introduction

This briefing note presents the key findings from a 2020[i] scoping study on the potential of combining Norwegian Church Aid (NCA[ii]) data with external datasets for better evaluations, and what doing that would require.

Through its work, NCA produces a significant amount of data for monitoring and evaluation purposes. This includes geo-located, household-level survey data as well as project locations (e.g. wells and other water supply points, schools). NCA also regularly conducts evaluations of a global nature (multi-country evaluations).

The evaluation type used by NCA depends on the evaluation's overall purpose and questions, the nature of the intervention to be evaluated, the existing evidence and available resources.

One area of interest is to explore the potential of combining datasets – data produced by NCA together with open databases. This includes:

- Traditional data sources (e.g. administrative data, survey data) made accessible with open data protocols, improving the possibilities for finding and reusing data files.
- Geographical data combining points (e.g. geolocated events), polygons (e.g. borders or water shores) and/or raster images (e.g. satellite photographs) that can be combined using powerful geographic information system (GIS) software.
- New data sources of increasing size, frequency and diversity (big data) collected *organically* by sensors (e.g. temperature or traffic measurements) or transaction machines (e.g. scanner or credit card data and mobile calls), with associated technologies for data storage and manipulation.

Evaluators' interest in big data has grown considerably in recent years, reaching a new peak as a result of the Covid-19 pandemic.

Requirements for datasets

In considering combining its data with open datasets, NCA has some requirements and preferences.

NCA's requirements include:

- Access to data files is granted with **minimum user access requirements**.
- Open data files are disseminated in **machine-readable formats** by widely available software.
- Data files are **accompanied by documentation** about methodological issues (i.e. metadata describing

variables, definitions and codes), quality parameters (geographical precision, sample size, reference date, etc.) and paradata about the collection process (time of data entry, identification of data entry operators, etc.).
- The data holder ensures **timely updates** of data.

Desirable characteristics include:

- **International coverage** (in the same file or in separate collection exercises), reducing the entry cost of understanding the data properties when using the information for different countries or interventions.
- Data files using **international geostatistical standards**[iii] **to define variables, classifications and breakdowns** and provide more opportunities for **inter-operability** with other files (i.e. comparing, linking and matching different datasets).[iv]
- Data content expressed as **coded variables**, minimising the use of open-ended questions and literal entries. Standard statistical techniques exist for treating numerical and categorical variables, while textual data analysis requires advanced machine learning.

Promising external datasets

Statistical sources (secondary sources)

Microdata from household or business sample surveys
The interoperability of household survey microdata and NCA data files is limited without advanced statistical techniques. However, there is high potential of using *statistical data matching* and *small area estimation* techniques with NCA household survey data, and could enrich NCA data (e.g. on gender-based violence or social cohesion) with other socio-economic variables.

Aggregated statistical data
There are two main limits to detailed NCA data: its representativeness (especially when collected via sampling) and the confidentiality of personal or sensitive information. NCA can use aggregated statistical data as part of needs assessments to identify areas and thematic priorities for interventions. However, such datasets have a limited potential for evaluation, given the difficulty of linking variables from external datasets and the impact of NCA interventions.

Open geocoded data

This might include geostatistical data files, or any other type of content with place-related information.

In addition to basic geographical layers such as administrative borders and locations, geospatial files provide information about natural and human-made environments (land use, bodies of water, obstacles, infrastructure, etc.) and events (e.g. violent outbreaks or natural disasters), including real-time data (e.g. meteorological data). Geospatial files have a high degree of international harmonisation and potential global coverage, since they are usually compiled by international agencies. Geospatial information has the advantage of easy visual interpretation, which facilitates dissemination to non-specialist users. Furthermore, file formats are increasingly usable by non-GIS software.

A special difficulty of geospatial data is the geographical coordinate system. Several standards will have to be mastered before linking NCA data and geospatial data files, as they need to use the same system to avoid representation issues.

Open administrative data

Administrative data might include lists of projects, budgets, etc. This type of data often includes textual variables, such as names, addresses and project titles, which cannot be directly subjected to statistical analysis. Textual data requires the formation of thesauri to carry out semantic analysis, or manual processing based on key word searches.

This means that high-potential administrative data for NCA interventions will need to be carefully selected. For instance, the International Aid Transparency Initiative (IATI)[v] datastore in certain locations seems to be relevant for NCA planning and evaluation, but would require preparation of a thesaurus relating to the thematic priorities (preparing a list of key words related to gender-based violence, water, sanitation and hygiene and peacebuilding, etc.) before IATA and NCA data could be usefully integrated.

Applicability in practice – an example from Somalia

NCA ran some experiments linking NCA datasets with pre-selected external sources to test the utility of the approach and answer evaluative questions, using an NCA gender-based violence (GBV) dataset from Somalia. This dataset comprised 200 individuals in nine internally displaced person (IDP) camps in two districts of Somalia (Garowe and Mogadishu), from 12–17 February 2019. It contains 108 variables, 12 of which are automatically recorded paradata.

NCA aimed to test whether it could be possible to address questions related to relevance (To what extent is NCA intervening in the areas of greatest GBV prevalence?) and coherence (To what extent is NCA's work consistent with the intervention of other actors?). External data from the following sources was used, considering recorded GBV incidents, humanitarian and development projects, and risks factors as proxies (e.g. lighting and distance to water sources):

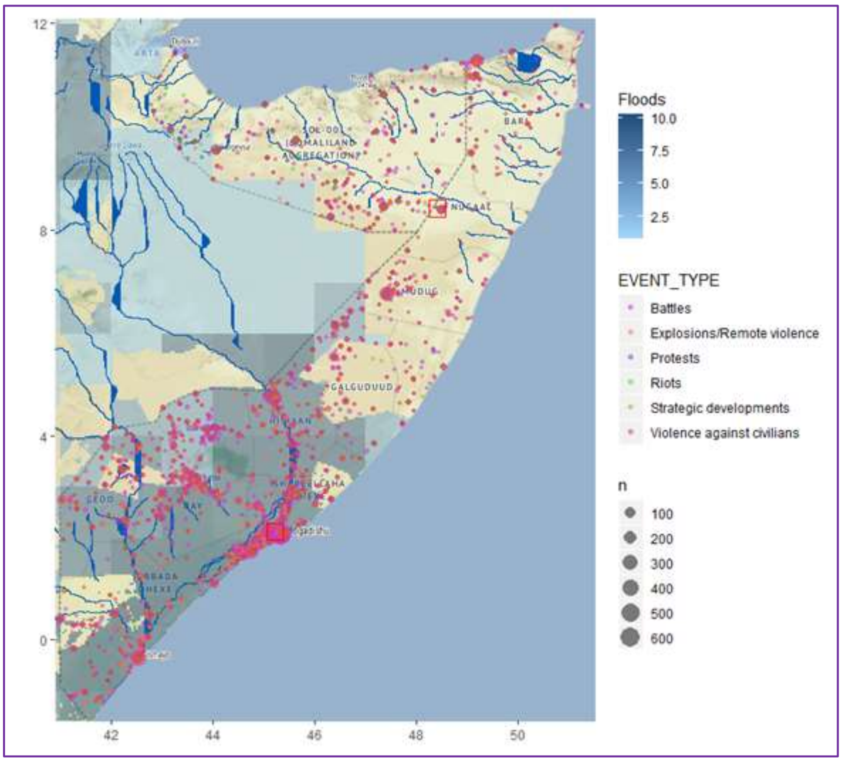

Figure 1 Combined information from ACLED, GSHHG and Global Flood Hazard Frequency and Distribution, Somalia. (NB: Large-scale GIS mapping plotting, e.g. access to safe houses, was also produced but is not included here due to sensitivities)

- GSHHG Distance to Water
- Armed Conflict Location & Event Data (ACLED)
- IATI Datastore
- Global Flood Hazard Frequency and Distribution
- Estimated number of IDPs at sites assessed by CCCM
- CCCM Cluster Somalia Detailed Site Assessment (DSA)

These analyses allowed to draw some lessons on the evaluation process, opportunities and limitations - and confirmed that NCA GBV interventions captured in that dataset from Somalia are externally relevant and coherent.

Some conclusions

The main conclusion from NCA's scoping study is clear: it is indeed possible for NCA to maximise the use of its internal data in evaluations by pairing or matching it with openly available datasets. This presents a wide range of possibilities to address OECD DAC evaluation criteria, especially in relation to relevance, coherence and impact.

Advanced GIS and spatial statistics can combine multiple data sources to model “catchment areas” for services, “risk areas” for violence or environmental hazards, and also to produce more complex metrics (e.g. calculating spatial correlation coefficients).

There are, however, a series of prerequisites in order for this to be effective. Properly merging multiple data sources may require a data analysis specialist who can manage huge amounts of data using powerful languages/programs such as R or/and Python, and advanced GIS and spatial statistics. Merging data files is challenging. Even when all sources use standard geographic variables, file formats and the way variables are stored can make it difficult to harmonise all sources. Furthermore, working with certain databases requires skills in the statistical analysis of textual data, when external sources include literal descriptions of interventions.

Evaluators and evaluation managers also need to become familiar with the basics of both the analysis and characteristics of internal and external databases. More importantly, a certain degree of evaluation ingenuity is a must; technical requirements and skillsets are not much help if NCA cannot imagine and craft the right evaluative questions, considering the wealth of data and its full possibilities. Technical and resource challenges in combining datasets can be overcome as long as the creative mindset is there, but it is hard to get something of worth without the right evaluative thinking and analytical framework.

[i] The scoping study was designed by Norwegian Church Aid and carried out by DevStat in January–June 2020. Contacts: Javier Fabra-Mata, Senior Advisor for Evaluations and Research, javier.fabra-mata@nca.no; Andrej Viotti, Methods, Evaluation and Learning Unit Team Leader, andrej.viotti@nca.no . Release date: 3 August 2020.

[ii] Norwegian Church Aid is a diaconal organisation mandated by churches and Christian organisations in Norway to work with people around the world to eradicate poverty and injustice. Thematically, NCA’s 2020–2030 development work comprises three global programmes and three strategic initiatives

[iii] International geostatistical standards include: geodetic system coordinates, geographical codes and place names; and defined socio-economic and environmental variables based on Sustainable Development Goal indicators, UN Recommendations for Population and Housing Censuses or sector-specific classifications (e.g. goods and services, economic sectors, occupation status, education levels, etc.).

[iv] Interoperable data can easily be reused and processed in different applications, allowing different information systems to work together. Interoperability is a key enabler for the development sector to become more data-driven.

[v] https://iatistandard.org/en/