Compilers' Manual on Forced Displacement Statistics

USE CASE

F

NON-TRADITIONAL DATA SOURCES

SOUTH SUDAN.
South Sudanese men fleeing Sudan violence try to find a phone signal at the Joda border point near Renk in 2023.
©UNHCR/Andrew McConnell

EGRISS

This is Use Case F from the *Compiler's Manual on Forced Displacement Statistics*. The Use Case describes the range of non-traditional data sources that may support the production of official statistics on displaced populations.

The *Compiler's Manual* is aimed primarily at technical personnel in National Statistical Systems who want to include displaced populations – refugees and / or Internally Displaced Persons (IDPs) – in official statistics. Each Use Case discusses a different scenario relevant to producing official statistics on refugees and IDPs, with a focus on the elements of statistical production cycles that are specific to refugee and IDP contexts. Spotlight examples of good practice in the production of refugee and IDP statistics are interwoven throughout the Use Case.

The Compiler's Manual and its individual Use Cases are intended to be a 'living document' which will be amended and extended as the body of expertise and knowledge develops worldwide.

Note: Paragraph numbering is per the complete version of the Compilers Manual.

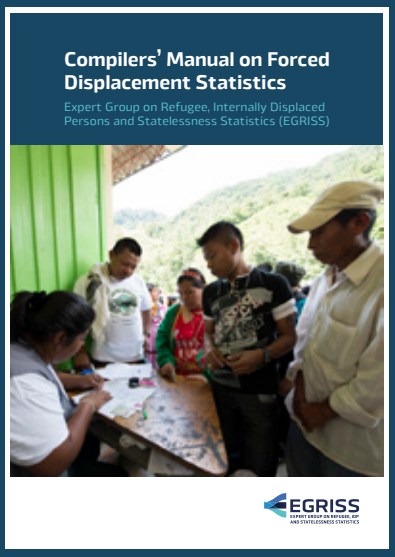
The Expert Group on Refugee, IDP and Statelessness Statistics

The Expert Group on Refugee, IDP and Statelessness Statistics is a UN Statistical Commission mandated, multi-stakeholder group that works with National Statistical Offices, international organizations and civil society to develop and support implementation of international standards and guidance to improve official statistics on forcibly displaced and stateless persons.

The Compilers' Manual

The Compilers' Manual offers clear operational instructions on producing official statistics on refugees, asylum seekers, IDPs and related populations. It complements the content of the International Recommendations on Refugee Statistics and the International Recommendations on Internally Displaced Persons Statistics by providing hands on guidance.

Access the complete version of the Compilers' Manual'



egrisstats@unhcr.org | https://egrisstats.org | twitter.com/EGRISStats

TABLE OF CONTENTS

Non-traditional data sources – general background

191. Non-traditional data sources encompass the huge volumes of data generated by new technologies such as mobile phones and social media. These provide a rich potential source to complement or aid the production of government official statistics, with the benefit that they are often more timely and detailed than traditional statistics. The UN has established a committee of experts on big data and data science for official statistics[1] that advocates for the use of non-traditional data sources to complement and improve national official statistics. The UN Statistics Division has stated that "big data constitute a source of information that cannot be ignored and that the global statistical community must organize itself and take urgent action to exploit the possibilities and harness the challenges effectively"[2].

192. The potential value of non-traditional data sources in relation to displaced people is significant, for example to fill gaps in understanding around the number and location of displaced people in rapidly evolving situations or to provide information that can improve sampling frames for surveys. However, it is important to be aware of the limitations and caveats that apply across most of these sources.

- The data are not produced for statistical purposes so there should be no expectation that data quality is sufficient for the production of official statistics. As with the use of operational data, the statistical authority in charge (e.g. NSO, line ministry, specialised statistical unit within NSS) will need to determine the usefulness of non-traditional data sources within official statistics production processes, understand where the data come from and assess their quality.
- Non-traditional data sources are often owned by private companies and legal use of the data will require negotiation of data sharing agreements, consideration of data protection and may potentially require a payment.
- Many of the sources rely on individuals interacting with technology, for example through a mobile phone, use of social media or the internet. This creates an inherent bias, particularly in relation to age (older age groups and the very young are unlikely to be represented) and access (excludes remote rural locations with limited connectivity) but also potentially in relation to gender, disability, income or other characteristics.
- Identification of refugees and IDPs through technology is not equivalent to identification following a set of survey questions, for example, and usually requires assumptions to be made. There can be important definitional differences which need to be considered, particularly if using non-traditional sources as a source of statistics for comparison or to augment survey results: the two may not be covering the same population, which will make comparisons invalid or problematic.

193. Provided these challenges can be overcome, the sources can hold a lot of promise, particularly in terms of complementing or supporting the production of traditional sources, if not yet as a replacement for them. This section focusses on providing a summary of the non-traditional data sources, what each is useful for, how to arrange access to the data, and provides case studies to illustrate and inspire.

1 United Nations. UNBigData

2 Big data and modernization of statistical systems

Mobile phone call detail records

194. *What is it?* When mobile phones are used to make or receive calls (or texts) the telecommunications provider that handles the call generates a record that includes many attributes of the call, such as the time, duration, the phone numbers making and receiving the call, and location details. This generates huge volumes of data, known as Call Detail Records (CDR) and owned by the telecommunications companies.

195. *What is it useful for?* CDRs can provide up to date information about sudden changes in population density in small areas, informing understanding of the scale of displacements and potentially the size of new refugee or IDP settlements. This can help, both directly with estimating stocks and flows of displaced people but also to support sampling and enumeration in a survey or census. With more advanced analysis CDRs can support understanding of refugee integration through analysis of call location patterns and calls within or outside the refugee community.

196. *How to access it?* CDRs are proprietary to the carrier network and cannot be accessed unless a bilateral agreement between the analyst and the carrier network is in place that addresses privacy concerns. This may involve limited release of a restricted data sample, sharing of non-identifiable aggregated or anonymized data, or remote access to anonymized data on a virtual environment controlled by the mobile phone operator. If the population movements of interest cross borders, the number of agreements needed – and the complexity of the challenge – rises and it can be more practical to look for other sources, such as internet-based social media records. The UN Global Working Group on Big Data for Official Statistics has published a *Handbook on the Use of Mobile Phone Data for Official Statistics*[3].

3 Handbook on the Use of Mobile Phone Data for Official Statistics

CASE STUDY: POPULATION DENSITY MAPS IN NEPAL

In April 2015, a devastating earthquake with a magnitude of 7.8 struck Nepal. Following the earthquake, Flowminder was given access to anonymised mobile phone data by Ncell, the largest mobile operator in Nepal. Flowminder accessed the data directly at the operator`s premises using it to produce static population density maps after the earthquake, which were used by UN OCHA and other key relief agencies to estimate the number of people displaced and affected, and coordinate disaster relief operations. The project mapping team estimated that around 1.8 million people above normal levels had left their home districts because of the 2015 earthquake.

Source: Flowminder (2015). Nepal Earthquake 2015 Case Study. Retrieved from https://web.flowminder.org/case-studies/nepal-earthquake-2015.

CASE STUDY: SEGREGATION OF SYRIAN REFUGEES IN TURKEY

Researchers have used CDRs to analyse patterns of spatial segregation of Syrian refugees in Turkey, and how patterns of spatial segregation influence internal mobility decisions of refugees as they move to other regions within the country. This enables the construction of two indices of integration: a dissimilarity index measuring the share of refugees that would have to move from high to low concentration regions to match their average distribution across the country, i.e. to achieve full integration; and a normalized isolation index measuring the probability that refugees interact with the wider population. The analysis is based on anonymized CDRs from Turk Telekom for their Syrian customers and for a large sample of Turkish customers. The database includes phone activity for a sample of nearly one million customers, out of which approximately 185,000 are tagged as refugees. The granularity of the data permits the analysis of calls made/received at a geographically disaggregated level (i.e. for each cell phone tower) for each hour in 2017.

Source: Segregation and internal mobility of Syrian refugees in Turkey: Evidence from mobile phone data. Retrieved from https://www.jointdatacenter.org/literature_review/segregation-of-syrian-refugees-in-turkey-evidence-from-mobile-phone-data/.

CASE STUDY: LEVERAGING BEHAVIOURAL AND HUMANITARIAN DATA SOURCES TO ANALYSE DEVELOPMENT CHALLENGES FACED BY SYRIAN REFUGEES AND HOST COMMUNITIES IN LEBANON

UN ESCWA, in partnership with the Qatar Computing Institute (QCRI) and the Data-Pop Alliance, explored the potential of mobile phone data to generate timely, granular, and cost-effective estimates for the vulnerabilities faced by refugees and host communities in Lebanon. CDRs from two mobile operators, Alfa and Touch, were analyzed through the Ministry of Telecommunications. Data from Touch consisted of the number of outgoing and incoming calls, disaggregated by age and gender; data from Alfa contained statistics about data consumption. The spatial distribution of calls proved to be a good indicator of population distribution, correlating closely with traditional official statistics. The ratio of dial-out-duration to receiving-in-duration in Touch data turned out to be a good predictor of socioeconomic factors, through correlation with the number of Syrian refugee families with debt greater than USD $600 and correlation with the number of self-declared poor in the Lebanese host community.

Source: Leveraging Behavioral and Humanitarian Data Sources to Analyze Development Challenges Faced by Syrian Refugees and Host Communities in Lebanon. Retrieved from https://archive.unescwa.org/publications/big-data-challenges-syrian-refugees-lebanon

Mobile phone GPS data

197. *What is it?* Most smartphones collect Global Positioning System (GPS) data, which is used by many apps to track the phone's location and tailor the service on offer accordingly. GPS data differs from location data available through CDRs: the latter rely on identifying which mobile phone tower / receiver has been used by the phone, so is less precise and the data are owned by the telecommunications provider. By contrast, GPS data uses satellites and is more accurate. The data tend to be collected (and therefore owned) either by the phone system manufacturer (for example, Apple) or by the companies that own the apps that use the GPS data. Users typically allow GPS location data to be collected by responding to notifications on their devices that ask whether they are willing to share their location data. The way this happens varies across apps and across phone brands.

198. *What is it useful for?* GPS data provides detailed accounts of movement patterns, for example to understand daily commutes; it can be used to ensure survey enumeration is occurring in the correct location, and the inclusion of GPS technology in smart phones and tablet computers can integrate navigation and the recording of geographical coordinates into enumeration activities, minimising locational error and human effort.

199. *How to access it?* Third party brokers can provide access to data at a cost. There are few examples of the use of GPS data that are directly relevant to producing statistics on displaced populations, but the possibility is included here for completeness. There are literature reviews available that illustrate how mobile phone data (both CDR and GPS) have been used for statistics, for example from the US Census Bureau[4] and the UK Office for National Statistics[5].

Satellite imagery

200. *What is it?* Satellites are used by governments and businesses to take detailed images of the earth's surface, which can build up a regularly updated photographic record of geographical features.

201. *What is it useful for?* Satellite imagery can be used to generate detailed geographical breakdowns, for example to show the locations of displaced communities, either to inform estimates of stocks or to help construct sampling frames. They can also provide geographic context such as proximity to nearby services, schools and hospitals.

202. *How to access it?* Access to some satellite imagery can be free, while many sources are available at a cost.

CASE STUDY: GENERATING ENUMERATION BLOCKS FROM SATELLITE IMAGERY FOR A SURVEY IN SOMALIA

In Somalia where security considerations prevented the usual sampling techniques to be carried out an innovative technique was used. It was not feasible to conduct a full listing of all households in an enumeration area, as this was too time-intensive and raised security concerns. Instead, enumeration areas were segmented into smaller enumeration blocks using satellite imagery. Enumeration blocks are small enough for enumerators to list and select households immediately before conducting the interview.

Source: Utz Pape and Philip Wollburg. "Estimation of Poverty in Somalia Using Innovative Methodologies." World Bank Policy Research Working Paper 8735, February 2019 https://documents1.worldbank.org/curated/en/509221549985694077/pdf/WPS8735.pdf

4 Use of Mobile Phone Location Data in Official Statistics, Social, Demographic and Health Studies

5 ONS methodology working paper series no. 8- Statistical uses for mobile phone data: literature review

CASE STUDY: USING SATELLITE IMAGERY TO ESTIMATE SIZE OF ENUMERATION AREAS IN SOUTH SUDAN IDP SURVEY

The Crisis Recovery Survey was conducted in 4 IDP camps in South Sudan between May to July 2017. The sample was restricted to Protection of Civilian (PoC) camps, and includes the 4 largest camps with clearly defined boundaries. The sample was designed as a multi-stage stratified random sample. Each camp was selected as a strata, with a target of 600 interviews per camp. Within each camp, 50 enumeration areas (EAs) were selected proportional to size, where the size was defined by the number of structures in the EA. The number of structures was estimated using satellite imagery of the strata (camps). Each EA was divided into 12 blocks, and a micro listing was done in the blocks to randomly select households.

Source: https://microdata.worldbank.org/index.php/catalog/2914

Social media data

203. *What is it?* Social media generates huge volumes of data, through the content of material that individuals post on social media, such as Facebook and Twitter, as well the locations and times of posts and characteristics of the individual posting.

204. *What is it useful for?* All the data has potential analytical value but applying them to inform refugee and IDP statistics is still new and has mostly happened in a research context rather than generating a set of readily repeatable analytical techniques. For example, sentiment analysis of social media posts (through analysing the text to understand positive or negative sentiments) can be used to support analysis of refugee integration in host communities, while geo-tagged tweets and Facebook data have potential to inform understanding of flows of IDPs and refugees. This can be particularly valuable in the context of cross-border flows, where mobile phone data are of limited value as mobile phone SIM cards are linked to national providers, so human mobility calculated from phone records can only be used to estimate movements within countries.

205. Currently there are few examples of social media data providing useful analysis to support official statistics on refugees and IDPs, but there is research that shows the potential, as illustrated in the following case studies. Research focussed on this topic mainly uses Facebook's advertising platform or covers sentiment analysis. The main limitation of using social media sentiment analysis is that the data can be biased by automated or spam tweets: the extent of social media posts published by automated "bots" is unknown.

CASE STUDY: MONITORING OF THE VENEZUELAN EXODUS THROUGH FACEBOOK'S ADVERTISING PLATFORM

Venezuela is going through the worst economical, political and social crisis in its modern history. Basic products like food or medicine are scarce and hyperinflation is combined with economic depression. This situation is creating an unprecedented refugee and migrant crisis in the region. Governments and international agencies have not been able to consistently leverage reliable information using traditional methods. Therefore, to organize and deploy any kind of humanitarian response, it is crucial to evaluate new methodologies to measure the number and location of Venezuelan refugees and migrants across Latin America. Research has proposed using Facebook's advertising platform as an additional data source for monitoring the ongoing crisis. National and sub-national numbers of refugees and migrants have been estimated and validated and socio-economic profiles have been disaggregated to further understand the complexity of the phenomenon. Although limitations exist, the presented methodology may be of value for real-time assessment of refugee and migrant crises world-wide. Similar techniques have been deployed in monitoring displacement caused by the war in Ukraine (see second link).

Source: Monitoring of the Venezuelan exodus through Facebook's advertising platform

Source: Nowcasting daily population displacement in Ukraine through social media advertising data

ECUADOR.
Venezuelan woman steps in to bridge digital gap in Ecuadorian neighbourhood in 2023.
©UNHCR/Santiago Arcos Veintimilla

CASE STUDY: REFUGEE CAMP POPULATION ESTIMATES USING AUTOMATED FEATURE EXTRACTION IN BANGLADESH

High-resolution satellite imagery can be used to map (manually or through automated feature extraction) physical structures in refugee and IDP camps, including changes to the number and type of these structures over time, to support population estimates and geospatial analysis. This technique was applied in a study during the Rohingya refugee crisis, focusing on areas in and around existing refugee communities in two main refugee settlements in Bangladesh. Population estimates for each of the refugee camps were determined by: (a) identifying building features; and then using these features to (b) estimate the camp population based on the total area of the building features and UNHCR 'covered area per person' statistics. Automated feature extraction greatly reduced the average processing time for each camp. However, the accuracy of automated feature extraction methods rely on well-defined classifier definition files (used to classify pixels or group of pixels into different roof types and non-building features based on their spectral, textual or spatial properties) and it is not straightforward to establish well-defined classifier definition files that are geographically and temporally transferable.

Source: Refugee Camp Population Estimates Using Automated Feature Extraction

BANGLADESH. *In 2022, a fire devastates Rohingya refugee settlements in Kutupalong.* ©UNHCR/Amos Halde

CASE STUDY: A FRAMEWORK FOR ESTIMATING MIGRANT STOCKS USING DIGITAL TRACES AND SURVEY DATA: AN APPLICATION IN THE UNITED KINGDOM

An accurate estimation of international migration is hampered by a lack of timely and comprehensive data, and by the use of different definitions and measures of migration in different countries. In an effort to address this situation, traditional data sources for the United Kingdom have been complemented with social media data to understand whether information from digital traces can help measure international migration. The Bayesian framework proposed is used to combine data from the Labour Force Survey (LFS) and the Facebook Advertising Platform to study the number of European migrants in the United Kingdom, with the aim of producing more accurate estimates of the numbers of European migrants. The overarching model is divided into a Theory-Based Model of migration and a Measurement Error Model. The quality of the LFS and Facebook data has been reviewed, paying particular attention to the biases of these sources. The results indicate visible yet uncertain differences between model estimates using the Bayesian framework and individual sources. Sensitivity analysis techniques are used to evaluate the quality of the model. The advantages and limitations of this approach, which can be applied in other contexts, are discussed. No one individual source can necessarily be trusted but combining them through modelling offers valuable insights.

Source: A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom

OpenStreetMap

206. *What is it?* OpenStreetMap (OSM) is a collaborative project to create a free editable geographic database of the world. It is built through crowdsourced geographic information and in some locations its detail and precision rival "authoritative" datasets from governments and commercial entities. The data from OSM is freely available.

207. *What is it useful for?* As a mapping tool, OSM will potentially be a useful geographic resource to aid sampling and enumeration and potentially as a source of data to feed into other estimates. Ultimately, OSM's usefulness depends on the extent to which it is complete and regularly updated in the locations of interest.

208. *How to access it?* OSM is freely available through its website.

CASE STUDY: EVALUATING THE UTILITY OF OPENSTREETMAP DATA FOR MONITORING SUSTAINABLE DEVELOPMENT GOAL PROGRESS IN REFUGEE SETTLEMENTS IN UGANDA

All available OSM data within 28 refugee settlements and 26 non-refugee settlements in Uganda was collected. The data represents physical features associated with dwellings, schools, clinics, latrines, etc., with metadata on feature creation date, date of most recent edit (version), and descriptive tags. The study created a data model linking 149 OSM features to 11 SDGs. Based on these SDG-OSM pairings, the study: (1) quantified the spatial distribution of SDG-relevant OSM data across and within settlements; (2) measured the chronology of creation and subsequent versions of SDG data, and (3) compared the spatial and temporal coverage of SDG data between refugee and non-refugee settlements. Despite many limitations of this approach, the study concludes that the widespread availability of OSM data make it a promising source of information on SDGs in refugee settlements, as well as in peri-urban informal settlements and internally displaced person (IDP) camps.

Source: Development after Displacement: Evaluating the Utility of OpenStreetMap Data for Monitoring Sustainable Development Goal Progress in Refugee Settlements

UGANDA.
Landscape of the Benet Community in Chemukula village in 2021.
©UNHCR/Esther Ruth Mbabazi