Big Data for Good

Can Big Data Illustrate the Challenges Facing Syrian Refugees in Lebanon?

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity Dignified Life

إدارة الإحصاء المركزي

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

DATA-POP ALLIANCE

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

VISION

ESCWA, an innovative catalyst for a stable, just and flourishing Arab region

MISSION

Committed to the 2030 Agenda, ESCWA's passionate team produces innovative knowledge, fosters regional consensus and delivers transformational policy advice.

Together, we work for a sustainable future for all.

Big Data for Good

Can Big Data Illustrate the Challenges Facing Syrian Refugees in Lebanon?

United Nations publication issued by ESCWA, United Nations House,
Riad El Solh Square, P.O. Box: 11-8575, Beirut, Lebanon.
Website: www.unescwa.org.

Acknowledgements

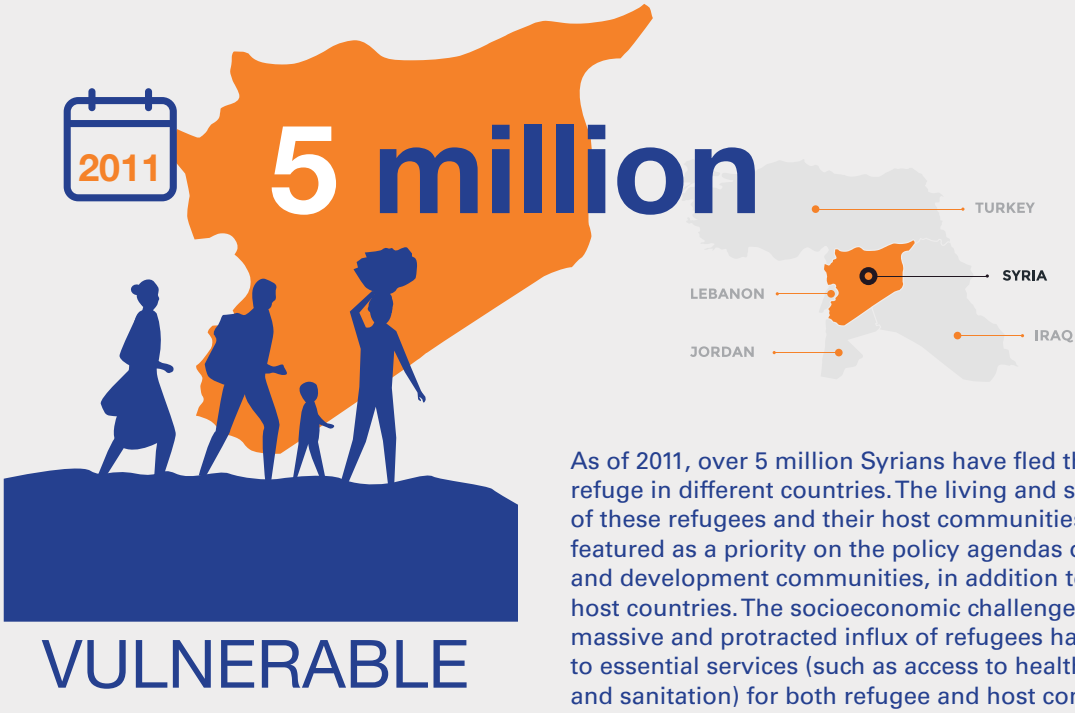
We would like to acknowledge the productive partnerships with the Data-Pop Alliance and the Lebanon Central Administration of Statistics (CAS). The contributions of our colleagues at the HBKU Qatar Computing Research Institute (QCRI) and UNHCR Lebanon are very much appreciated. On behalf of all the partners, we wish to express our appreciation for the Ministry of Communications in Lebanon for making the Mobile Call Detail Records available for the project team through CAS as custodian of data.

The management of data access, processing and storage was guided by clear principles agreed to with the multi-disciplinary CODE (Council for the Orientation for Development and Ethics) members.

Table of contents

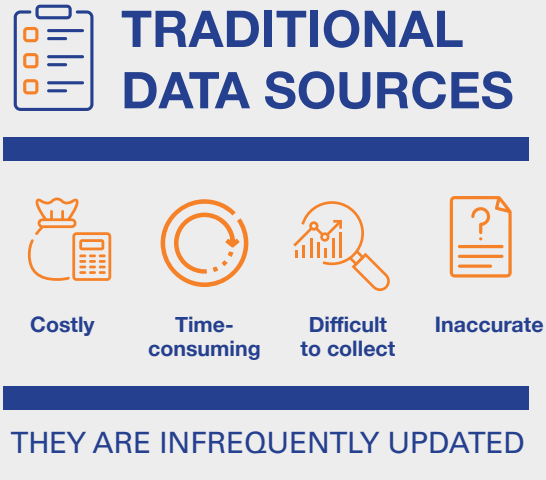
Key Messages

Introduction



As of 2011, over 5 million Syrians have fled their native land, seeking refuge in different countries. The living and social conditions of these refugees and their host communities have recurrently featured as a priority on the policy agendas of the humanitarian and development communities, in addition to decision makers in host countries. The socioeconomic challenges resulting from the massive and protracted influx of refugees have overwhelmed access to essential services (such as access to health care, schooling, water and sanitation) for both refugee and host communities, which in turn have further increased vulnerabilities and worsened conditions. These vulnerabilities are further aggravated by other shocks such as the pandemic and the resulting socioeconomic hardships on host countries.

In facing these challenges, policymakers and development and humanitarian practitioners rely on traditional data sources from official sources to inform their decisions with regards to both host communities as well as Syrian refugees. These sources, such as censuses and surveys, pose a few problems in this context. They can be costly, time-consuming and difficult to collect, and might be inaccurate. Also, they are infrequently updated, which can make the information they contain obsolete, especially within the highly volatile political and socioeconomic context of the Arab region. Policymakers must be able to offer timely and quick responses to the fast-paced and evolving challenges and crises experienced by refugees and their hosts.



A question then arises: are there other data sources that could be utilised by policymakers in crisis conditions? With the advent of the 'Data Revolution', we are currently surrounded by a wealth of information, known as "Big Data". Vast amounts of data are examined in order to extract patterns and insights, and recommendations and solutions are then formulated accordingly. Recently, policymakers have started to appreciate the potential use of big data to complement traditional sources of data for policymaking. This pilot project explores the potential of unconventional data sources in informing policymakers of the conditions and vulnerabilities of Syrian refugees and Lebanese host communities.

Findings

Demographic Status

Facebook data indicates that there are 2 to 3 times more male users registered than female users (2019).

Phone calling activities suggest that the male population is around 6 times larger than the female population. Official statistics have males and females being on par demographically in 2018. The explanation for this discrepancy is that men are handling the registration of most phone lines for female family members.

CULTURAL FACTORS SUPPORT THE INTERPRETATION OF THE ABOVE DATA.

Economic Status

Facebook registration suggests that Syrian refugees and host communities have similar high school attainment, with a gap of less than 1 per cent. At higher education levels, the host community has 5 per cent higher attainment than refugees (in 2019). Despite the relatively small gap in educational attainment, refugees are much worse off economically than the host population.

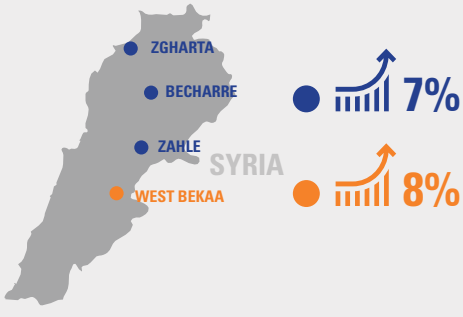
The annual (2016-2019) average duration of calls in Tripoli is approximately 85 per cent of the average duration of calls in Akkar (Tripoli's internet consumption is 97 per cent of Akkar's). Syrian refugees' activities display informal work behaviour (weekday and weekend calling activity do not differ much). This supports the belief that Syrian refugees are participating in informal work (such as farming) in Akkar.

Social Status

Scraping social media platforms and news articles revealed an increase in hostility towards issues related to Syrian refugees (negative rating increased by 9 per cent between 2017 and 2019 in a sample of 1054 Arabic articles on refugee-related incidents). Local Community tensions are growing steadily and with time these tensions can become dangerous.

The number of articles in Arabic covering Syrian refugees fell by 49 per cent between 2016 and 2019. On the other hand, the number of articles in English covering the same topic increased by 86 per cent.

Security & Mobility Patterns



As deduced from mobile call activities, the population of the municipalities of Becharre, Zahle and Zgharta increased by approximately 7 per cent during 2017, which could be due to the flow of Syrian refugees into these municipalities. Following an incident when a Lebanese woman was raped and murdered by a Syrian refugee in 2017, local population decreased by 8%. At that time, the population instead began increasing in West Bekaa, a region much closer to the Syrian border (by around 8 per cent).

Amid the tensions in Hermel in 2017 and 2018, the population of Hermel decreased by 1 per cent as the Baalbek population gained 1 per cent of the total Bekaa population based on calling activities. Community tensions led vulnerable populations to first relocate to neighbouring areas before considering areas farther away.

The proportion of Syrian refugees in the North and Bekaa decreased annually by 2 per cent (based on 2017-2019 Tawasol mobile calling data). This decrease could be attributed to the return of refugees to Syria and coincides with an increase of 80 per cent in English articles covering the return of Syrian refugees to their homeland.

Heavily populated areas with a district population of around 40 per cent or more of the total governate, as estimated based on call detail records, experienced an average of 6 per cent of Syrian refugees moving out of the area due to community tensions or restrictive measures, as reported by UNHCR VASyr 2018. A possible interpretation is that the concentration of refugees is correlated to security.

Recommendations

Technical Recommendations

Non-traditional data sources ought to be considered for supporting policymaking. In addition to official statistics or traditional data sources, the non-traditional data used in this project are useful in creating measurements of the well-being of Syrian refugees and their host communities. This is especially true for time-sensitive emergencies, where traditional data sources might be outdated or unavailable. In these situations, these alternative streams of information can assist policymakers in making decisions and "near real time" changes.

Researchers and Telecom Operators should maintain close communication, in order to pave the way for fruitful research and action to exploit the potential of calls and internet consumption data, all while maintaining the privacy of users.

Policy implications

Apart from open sources, policymakers should take advantage of the wealth of data that are only available to them.[1] The digital footprint left by users' activities can prove to be of great usefulness in estimating various indicators of population conditions[2].

Host communities should be informed of the educational attainment and economic potential of refugees residing in their area. This can allow better social and economic integration[3] for the interim period until refugees return home.

Refugee communities, through international organizations, NGOs and partnering associations, should maintain open communications with host communities and local officials in order to mediate any potential conflicts and control any security threats to either side.[4]

1. For example, only the Ministry of Telecommunications and the Telecom operators are privy to the calls data in this study.

2. We were able to obtain indicators on the economic and social condition of refugees and host communities from this data, which correlate with official sources of statistics.

3. This can be evidenced by the pay gap between host and refugee communities despite the small difference in education level, or the informal work by refugees which was picked up by the phone calls data.

4 The mass eviction from the whole region in the North, which coincided with one incident in Zgharta, is a clear indicator of the need for better synchronization between leaders of refugee and host communities.

01 Introduction

Timely evidence for policymaking has been a consistent objective for UN organizations and member states. The classical data sources needed for official statistics have been expensive, time consuming, leading to delayed outcomes, and in certain cases impossible to access (e.g. images, videos). Today's connected world offers alternatives and supplementary non-traditional data sources enabled by Information and Communication Technology (Geographic Information System, call data records, social media, open data, others).

The Arab countries members of ESCWA have been evaluating these sources for official statistics and possible public policymaking. Specifically: ESCWA Report of the Thirteenth session of the Statistical Committee Beirut, 29-30 January 2019: Recommendation (e) Continue updating and developing national statistical systems, benefit from technology in producing, collecting and using information and data, and focus on geospatial technologies and new technologies in censuses and surveys, while taking into account that the success of such processes relies on the availability of technical and financial support, especially in evaluating the infrastructure and readiness of statistical offices.

In this context, this pilot project aims to explore the potential of non-traditional data sources to produce information on a range of indicators – including sociodemographic characteristics – useful for policy creation, design and planning for Syrian Refugees and Host Communities in Lebanon (Figure 1). More specifically, the study analyses the predictive capacity of non-traditional data sources, obtained through secure and privacy preserving processes, to provide insights on the development challenges faced by refugees and their host communities.

Multiple efforts have been carried out to understand the vulnerabilities and challenges faced by refugees and host communities in Lebanon. At the same time, more timely, granular and cost-effective estimates remain an important priority.[5] With the constant movement and resettlement of forcibly displaced populations and refugees, timely and accurate estimates of where and how this population is interacting with host communities are crucial for the design of policies. These insights are crucial for planning systems and services available to host communities and residents, as well as for developing policies aimed at decreasing the vulnerability, risks and challenges faced by both populations.

With the constant movement and resettlement of forcibly displaced populations and refugees, timely and accurate estimates of where and how this population is interacting with host communities are crucial for the design of policies.

This backdrop comes in the context of the 'Data Revolution' of the past decade, which has created expectations – backed by a rich body of literature and some, albeit still slim, hard evidence – regarding the potential of so-called 'non-traditional' data sources (described below) and applications of computational methodologies to address complex human challenges. Yet, despite efforts over the past few years to standardise and systematise the use of these non-traditional data sources, more work remains to overcome existing challenges, including privacy concerns, data sharing hurdles, legal and regulatory barriers, among others.

As an exploratory analysis, a series of non-traditional data sources were chosen, to explore and analyse their potential to affirm new insights or confirm previously calculated estimates. Through the identification of various data sources and the assessment of data quality, prospective usefulness and applications, feasibility – in access, processing and analysis – and accuracy, the project sheds light on how different data sources independently as well as collectively can yield useful information for the analysis of the conditions and characteristics of host communities and refugee populations.

This project mainly explores the potential of. (1) Call Detail Records (CDRs), (2) Facebook Advertising Data, (3) the Global Database of Events Language and Tone (the GDELT project) and (4) Twitter data, for this country-specific analysis.

The document is structured as follows: first, all the data sources used in this project are presented, followed by the methodology used in the analysis. Next, a section on data sources and approaches provides information on each specific data source, and how it was employed to proxy the relevant indicators for this project. Then, the predictive capacity of each indicator is tested with data from official statistics. Finally, the discussion and lessons learned sections conclude the document reflecting on the usefulness of the proof of concept.

Figure 1. Project expected outcomes

OUTCOMES-UP-TO-DATE STATISTICS

Population pyramid
(age, sex structure)

Refugees distribution
(space, heat maps)

Mobility patterns
(movement in time and space)

Border movement
(frequent& intensity of entry/exit cycles)

Economic activities
(branches of economy, formal vs infromal)

Health status
(access to health services, prevailing health threats)

School enrolment
(Attendance, dropout, completion by levels, territorial distribution)

Security
(reported and observed incidents, predictive measures)

Brief Presentation of the Data Sources

A. Non-traditional data sources: Big data sources

1. Call detail records

Call Detail Records are the traces that Mobile Network Operators record every time mobile phones are used. While recorded for billing purposes, these digital traces have been used over the past years as a critical data source to analyse patterns in human behaviour, including human mobility. Indeed, the high penetration of mobile phones and constant use throughout the day, allows CDRs to become 'behavioural footprints' of mobile users. In this sense, CDRs can be used to understand social and mobility behaviours of users or groups. These insights can include how many places a user visited in one day, or for example, the number of outgoing calls from a district. Through the application of algorithms and machine learning tools, researchers can identify geographical patterns and temporal trends in the use of mobile phones. It is worth noting that in order to preserve privacy, most of these descriptors and characteristics are computed at an aggregated level, so as to prevent the re-identification of individual users.

CDRs have already been used to examine refugee dynamics, in particular through the Data4Refugees (D4R) challenge, co-organised by Türk Telekom, Bogazici University and Tübitak in collaboration with Fondazione Bruno Kessler (FBK), MIT Media Lab, Data-Pop Alliance, UNHCR, IOM and UNICEF.[6] As a landmark data challenge, the D4R provided researchers across the globe with access to CDR datasets of Türk Telekom customers, including coarse and fine-grained mobility data, with the purpose of creating research on the health, education, unemployment, safety and social integration of refugees in Turkey.

The challenge's aims included contributing to the welfare of refugee populations and to gain information on key issues faced by these populations. It produced important insights, including methodologies and evidence on the impact of integration on the rate of measles infections and differences in communication patterns between border and non-border communities. Overall, CDRs have been applied in the past to characterise socioeconomic and demographic information of a population, to analyse behaviours, estimate inequality, estimate the impact of different events, and perhaps most recently, to assess the effects of Covid-19 on human mobility, among others.

2. Facebook advertising data

Facebook is the most used social media platform in Lebanon, with a share of almost 80 per cent of all usage of social media platforms.[7] One of Facebook's main streams of revenue is online advertising. For this reason, to attract more advertisements and improve the accuracy of audience targeting, Facebook has developed a targeted advertising platform, called Adverts Manager, which allows

Mobile CDRs have been applied to characterize socioeconomic and demographic information of a population, estimate the impact of different events, and most recently, to assess the effects of Covid19- on human mobility, among others. While Facebook is the most used social media platform in Lebanon, with a share of almost 80 per cent of all usage of social media platforms. It allows for targeting of dimensions that range from information explicitly reported by Facebook users such as gender, education, current location and home location, to information automatically inferred from their interactions on Facebook and affiliate websites.

advertisers to give detailed specifications of the type of users to be targeted with ads.[8] Using the Adverts manager, Facebook allows for targeting of dimensions that range from information explicitly reported by Facebook users such as gender, education, current location and home location, to information automatically inferred from their interactions on Facebook and affiliate websites. Facebook also allows marketers to connect with expatriates living within a given country, for example people born in Brazil living abroad, or expatriates living in a specific country.[9]

Information on how many Facebook users match certain criteria, such as currently living abroad, can be obtained free of charge through an appropriate API.[10] This type of data has been used to monitor the flow of refugees and migrants out of Venezuela[11] and to map poverty[12] and track digital gender gaps.[13]

3. GDELT

Supported by Google Jigsaw, the Global Database of Events Language and Tone (GDELT Project) is an open source near-real-time database that monitors the world's broadcast, print, and web news from all over the world in over 100 languages. GDELT's archive is as old as 1 January 1970. It updates every 15 minutes to collect near-real-time news by monitoring a vast array of sources; hundreds of thousands of global media outlets, 215 years of digitised books and 21 billion words of academic literature spanning 70 years.

Using deep learning techniques, GDELT identifies the people, locations, organizations, and other data about each article collected. In addition to the events database, GDELT provides another database called the Global Entity Graph (GEG) which calculates and stores the sentiment of most of the articles collected in the GDELT events database, also available in near real time since July 2016. The sentiment is annotated using Google's Cloud Natural Language API. Each article contains three sentiment values: polarity, magnitude and score, all of which describe sentiment for more than 17 languages. As of the time of writing this report, Arabic is not supported by the sentiment analysis API. The database is open source and allows anyone to either download chunks of the archive or access the database using Google Cloud Platform's BigQuery.[14]

Global Database of Events Language and Tone (GDELT) monitors the world's broadcast, print, and web news from all over the world, identifies the people, locations, organizations, and other data about each article collected. It calculates and stores the sentiment of most of the articles collected, also available in near real time since July 2016.

4. Twitter

Although Twitter is not the most popular social media platform in Lebanon, it is still used by many to express their opinions about multiple issues relevant to the country, with many users keeping their tweets public. Furthermore, Twitter provides a number of features for developers at three tiers: free, premium, and enterprise. The features of these tiers allow developers to automate many actions on Twitter, such as searching tweets, monitoring an account's activity, real-time tweet streaming, advertising on Twitter and automating direct messages.[15] In addition to the APIs created by Twitter and specially made for developers, multiple open source packages[16] are available that allow scraping for tweets based on specific keywords, topics, locations and hashtags.

B. Additional data sources: Ground-truthing

As opposed to traditional data sources, where data are usually collected through probability sampling and a measure of representativeness of the populations can be obtained, non-traditional data sources are collected without a specific probabilistic design. As a result, insights derived from these data sources are inherently biased, which poses challenges for scientific and policy uses. However, non-traditional data sources have the advantage of being more finely grained as well as collected at a higher frequency than traditional data sources. As such, combining both data sources can allow for more representative insights that can be updated faster and to finer levels of temporal and geographic granularity. For this purpose, indicators derived from traditional data sources are compared and tested against their proxies from non-traditional data sources, in order to test the

correctness of the approaches. In particular, we explore the following traditional survey data sources and compare their results with our calculated indicators:

1. The Vulnerability Assessment of Syrian Refugees in Lebanon (VASyR): this annual report has been published by UNHCR since 2013. Through a representative sample of Syrian refugee households, VASyR provides an overview of the geographical variations in vulnerabilities at the district and governorate levels.[17]

2. The Central Administration of Statistics Labour Force and Household Living Conditions Survey (CASLFS): this survey was conducted by the Lebanese Central Administration of Statistics (CAS) between 2018 and 2019 and was entirely funded by the Delegation of the European Union to Lebanon, with the technical cooperation of the International Labour Organization (ILO), Regional Office for Arab States.[18]

Methodology

The correlation and predictive power of different non-traditional data sources are tested against corresponding ground truth data from UNHCR's annual VASyR, and Official Central Agency for Statistics Labor Force CASLFS Survey 2019.

Table 1. Target and non-traditional variables explored in this work

	CDRs	Facebook MAU	GDELT
CASLFS – District-level population	X		
CASLFS – percentage of families self-described as rich, poor or average	X	X	
CASLFS – Labour force participation rate	X	X	
VASyR – percentage of Syrian youth (15-24) who are not in education, employment, or training (NEET)	X	X	
VASyR – (Survival) Minimum Expenditure Basket (MEB/SMEB) categories, percentage of below and above poverty line, household debt category	X	X	
VASyR – Out of families that recently moved, percentage of whose reason was threat or harassment, tension with community or restrictive measures, tension with landlord, etc.	X		X

In this section, we discuss the methodology used to assess the capacity of non-traditional data sources to provide useful insights and predict socioeconomic and demographic indicators normally through traditional methods. To this end, different proxies for socioeconomic or demographic characteristics are created, which take into account their relationship with the ground-truth indicator of interest. For instance, the number of calls made in a specific location can be related to the number of people residing there. In this case, the relationship between the number of calls and population proportions is tested. Following the same logic, the correlation and predictive power of different non-traditional data sources are tested against corresponding ground truth data from UNHCR's annual VASyR, and CASLF 2019. The following is the methodology followed for each of the non-traditional data sources we have. For some indicators, such as population mobility, there is no ground truth data to compare with. In these cases, the corresponding indicators are analysed with regards to their potential meaning in an indicative manner.

The indicators calculated in each of our approaches were calculated at the district level. Data was gathered for two governorates, Bekaa and North, meaning that we have a total of 12 districts in our dataset. In addition, we extracted relevant ground-truth target variables from VASyR's open source VAULT and CASLFS's open source annex data files. These target variables include socio-demographics, employment, incidents, relation with host communities, gender, age, and population distribution. VASyR's open source VAULT is available for the years 2018 and 2019, while CASLFS was based on data collected in 2018. Therefore, we have a total of 24 data points to be compared to VASyR's ground truth labels and 12 data points to be compared with CASLFS's ground truth labels. Table 1 shows the target variables extracted from CASLFS and VASyR and the non-traditional data sources that were used for the regression model and for the correlation analysis for these variables.

We first calculated the Pearson correlation coefficient between each of the calculated indicators and the target variables. The Pearson correlation coefficient measures the linear correlation between two variables X and Y. It ranges between -1 (total negative correlation) and +1 (total positive correlation), with 0 indicating the absence of any linear correlation.

We then performed a linear regression between each indicator and the relevant target variable, in order to establish whether a linear association exists. A Linear regression is used to model the relationship between a dependent variable (target variable from the traditional data sources in our case) and an independent variable (our calculated indicators from the different non-traditional data sources). In order to test the predictive power of our indicators in explaining our target variable, we fit a linear regression model for each district-level indicator. We also performed cross-validation and reported the out of sample R^2 score for each linear regression model. We perform multiple regression between combinations of indicators and the ground-truth label. For each target variable, we performed a K-feature selection, which selects the K best features from the relevant calculated indicators that predict the target variable based on a cross-validated R^2 score. We then calculate the feature importance of each of the selected K best features using a Decision Tree Regressor model that calculates the importance of each feature in predicting the target variable.

We model the relationship between the target variable from the traditional data sources and our calculated indicator from the different non-traditional data sources.

Data Sources and Approaches

A. Call detail records

As described above, Call Detail Records are the digital traces of phone usage that Mobile Network Operators record for billing purposes. While their main use is commercial, they have been used over the last several years as a critical data source to analyse human behaviours. A single call detail record contains various metadata and details about an individual call, such as time and duration of the call, source and destination numbers, caller and callee tower coordinates.

1. Description of uses in the literature

Many works have used call detail records to augment traditional statistical methods in different countries. Examples include predicting sociodemographic indicators such as population density and economic indicators such as poverty levels. In this section, we look at some of the relevant work that has been done in other countries using CDRs.

In the paper *Predicting poverty and wealth from mobile phone metadata*,[19] the authors use anonymized data from mobile phone networks to predict the poverty and wealth of subscribers on an anonymised individual level. They argue that CDRs can be a good reflection of numerous socioeconomic factors due to their ability to capture rich information such as pattern of travel, location choice, and histories of consumption and expenditures. The authors add that regionally aggregated CDRs have also been shown to correlate with regionally aggregated population statistics from census and household surveys.

The authors of *On the Relationship Between Socio-Economic Factors and Cell Phone Usage* [20] associate geographical units with their CDRs and the relative census data. Using statistical tests, they then determine the correlation between cell phone use variables and census variables. The census variables they determine range from demographic variables, such as gender, age and per cent population distribution in each geographical area, to educational and socioeconomic variables such as literacy rates, school enrolment, etc. The authors plot histograms representing the percentages of genders and age groups in addition to the socioeconomic levels of the census data as per the census and the CDRs. Both histograms exhibit similar distribution between CDRs and census variables. The authors also calculate mobility variables based on phone calls of users with their different places visited, distance travelled and the radius of travel.

The *Handbook on the use of Mobile Phone data for official statistics* [21] presents a survey of the usage of CDRs in multiple countries for socioeconomic, tourism, mobility, demographic, and educational purposes. In the paper *Overview of the sources and challenges of mobile positioning data for statistics*,[22] the authors propose several demographic indicators that can be generated using CDRs. These vary from the geographic distribution of population, internal migration and cross-border migration, short-term population trends (hourly, daily, weekly, monthly), real-time assessment for specific locations in case of gathering, emergency situations, disaster, and law enforcement to assess the impact of a situation and the support needed.

The authors of "*Dynamic population mapping using mobile phone data*"[23] use a dataset consisting of more than 1 billion CDRs from Portugal and France to estimate the population density at a national and seasonal scale. They show that not only can a population map be constructed via CDR population estimation, but this type of estimation based on daily phone records can show population dynamics.

In the paper "*Use of mobile phone data to estimate mobility flows*",[24] the authors investigate the usage of CDRs to detect inter-city population mobility and residence. Their results show that flow estimates in their study were more accurate for larger towns. In the paper "*Inferring Patterns of Internal Migration from Mobile Phone Call Records*",[25] the authors argue that mobile phones can be a new source for internal migration estimates. Their argument is based on their work with CDRs in Rwanda across four years and their ability to detect more subtle patterns that were not detected in the government survey.

Further works have examined Syrian refugees in Turkey. The authors of *Measuring fine-grained multidimensional integration using mobile phone metadata* [26] resort to CDRs to detect the social integration between refugees and host communities, the spatial integration of Syrian refugees based on the locations where they make their calls and the economic integration of refugees based on an employment score, which is based on the hourly number of phone calls of the refugees. The authors were able to differentiate between calls of refugees and calls of Turkish citizens and others based on the anonymised registration of the phone numbers.

2. Use for this project

For our project, we analysed CDRs to estimate different indicators; namely the population pyramid, population distribution, mobility patterns and economic activity of Syrian refugees and host communities in Lebanon. CDRs were requested in an official letter sent from the Lebanese Central Administration of Statistics (CAS), also a partner in this project, to the Ministry of Telecommunications in Lebanon. We received access to CDRs for the Bekaa and North Governorates from the two telecommunications companies in Lebanon, Alfa and Touch, for April, May and June of 2016 to 2019. In order to preserve user privacy, we only requested monthly site and cell-level aggregates of the CDRs, which summarised user calls at the site and cell-level without any details about any users. The data were stored on a machine at CAS that is not connected to the Internet and is only accessed by trusted personnel to calculate the indicators, plot variation graphs and perform other analyses.

(a) Touch data structure

Touch has a dedicated product for Syrians living in Lebanon called Al Tawasol. In our analysis, data on users of this bundle serve as a proxy for Syrian Refugees. For this purpose, the data received from Touch differentiated between the calls made by Touch users with regular bundles and those made by Al Tawasol users. This line "includes international minutes and SMS to Syria in addition to local minutes, SMS and internet, all for only $11 per month". The line includes 40 minutes for calls and 30 SMSs to Syria.[27]

For April, May and June of 2016 to 2019, the data obtained from Touch contained:

1. Yearly cell-level aggregates for incoming and outgoing calls.
2. Gender and age based yearly aggregates for calls.
3. Daily aggregates of the average number of calls per hour made by Al Tawasol users (2017-2019).

(b) Alfa data structure

Since Alfa does not have a special line for Syrian users, the data received was used to proxy the information of the broader category of Lebanese residents and other nationalities. The following was provided by Alfa:

1. Monthly aggregates of incoming and outgoing calls at a site level with governorate and district details.
2. Download, upload and total mobile data consumption in MB per district.

It is worth noting that the registrations of lines with both telecom companies are not fully accurate representations of individual behaviour, since phone ownership can change without the operators' knowledge. Examples include exchanging phones between family members, and not always registering phones (sim cards) under the user's name upon purchase. Additionally, the underlying assumption that Al Tawasol users can capture Syrian refugees and that other bundle users from Touch and Alfa represent host communities possesses its own limitations since Syrian refugees may purchase other bundles. For this reason, testing the predictive capacity of these indicators with appropriate ground truth data is a fundamental part of this proof of concept.

3. Results

The following section presents the indicators that were calculated using CDRs. The objective was to provide insights into population distributions, mobility and economic activity through simple indicators that could be easily replicated in future contexts. As an exploratory and indicative example of analysis in the context of a crisis, where indicators generated in near real-time can provide the most valuable insights for decision making, some of these results are put into context of security-related events that affected the refugee population.

(a) Population distribution

The dataset consists of the CDRs of the Bekaa and North Governorates. To approximate population distributions, site-level calling records are aggregated to the district level. Then, the proportion of calls by district in each governorate is computed to serve as surrogates for the distribution of population. The underlying logic is that average calling behaviour among individuals would not systematically differ when aggregated across a large space. Additionally, the number of mobile phone lines in a given administrative region is held to be proportional to the number of residents. Put simply, more calls are expected in areas with more people and vice-versa. The indicator is calculated as follows:

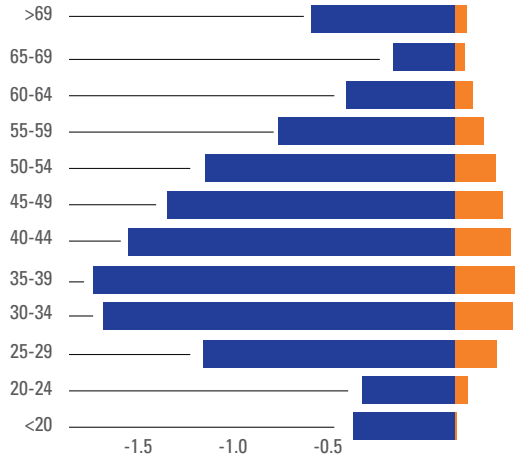
$$\textit{district population proportion} = \frac{\textit{Total number of calls at all sites in District}}{\textit{Total number of calls at all sites in Governorate}}$$

For April, May and June of 2016 to 2019, the data contained Yearly cell-level aggregates for incoming and outgoing calls; Daily aggregates of the average number of calls per hour; Download, upload and total mobile data consumption in MB per district. Mobile Operator Touch has a dedicated product for Syrians living in Lebanon called Al Tawasol.

(b) Population pyramid (gender and age distribution)

In order to explore the potential of CDRs to reveal gender and age distributions, the demographic information of Touch users is plotted. This information corresponds to the aggregated age and gender of callers.[28] Figure 2 shows the Touch-based population pyramid for the year 2018 with number of calls in ten million (along the x-axis, with males to the left to form a pyramid). The graph is highly skewed towards males aged between 30 and 55 years of age. This behaviour is most likely due to a single individual acquiring different lines and then selling them or giving them to other users that go unregistered. To this extent, this bias towards males is one of the limitations of the CDR data in this context.

Figure 2. TCDR-based population pyramid (2018)



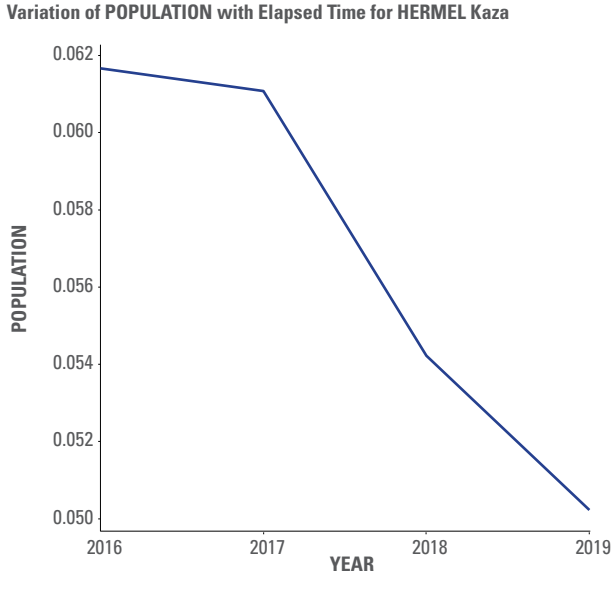
(c) Population mobility

Analysing the changes of population proportions across time can help creating proxies for mobility. Using the population indicator calculated above, the district-level net mobility is presented across different years, based on the change of proxied residents in each district.

(i) Variation in annual population

Figure 3 displays the variation in the population proxy for Hermel District. A continuous decrease in the calling behaviour is observed, which could suggest changes in population distributions. The downward trend between 2016 and 2017 is followed by a more pronounced and accelerated decrease in the population proportion (our proxy indicator) between 2017 and 2018, with some attenuation after 2018. It should be noted that in 2018, where the Lebanese Army carried out large-scale drug raids in the Hermel District.

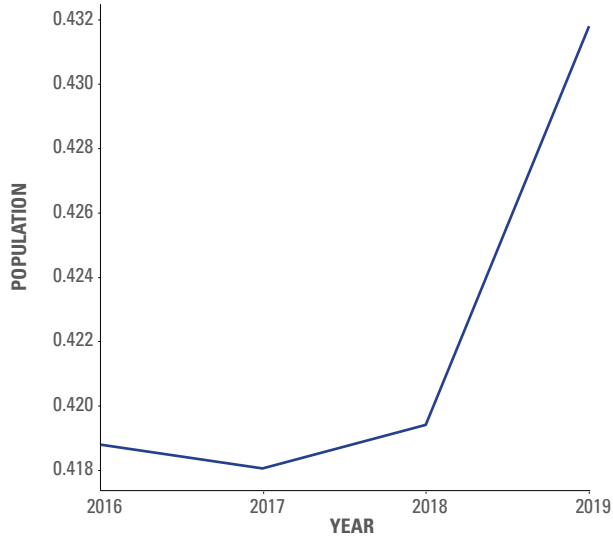
Figure 3. Population of Hermel District



(ii) Extracting trends from annual population variation

Furthermore, calling patterns can present seasonal variations and/or trends, which could lead to noise that confounds changes in users and population. For this reason, we inspect the population's time series as displayed in Figure 4. This plot shows the variation in the population of the Baalbek District across years, for the months selected in our analysis.

Figure 4. Population proportion of Baalbek



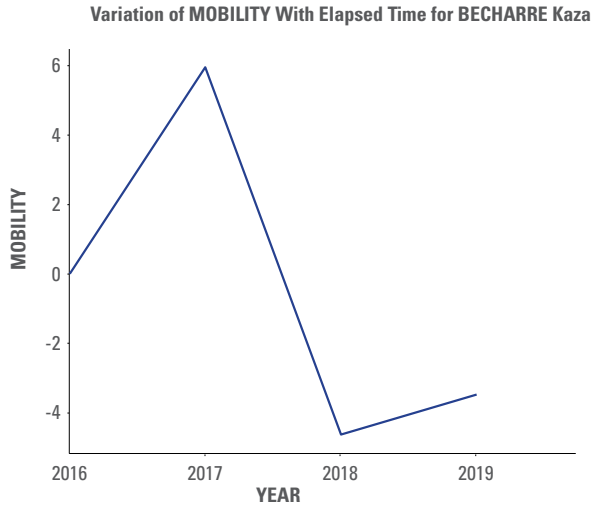
(iii) Annual mobility indicator

Taken as an index, percentage changes across time can suggest mobility changes. For this analysis, district-level mobility is presented as the percentage net variation of the population proxy indicator using the following formula:

$$annual\_mobility = \frac{population\ (current\text{-}previous)\ year}{population\_previous\_year} \times$$

Figure 5 shows an example of the variation in the annual mobility indicator for Becharre District. The population increased by 5 per cent in 2017 and decreased by 5 per cent in 2018 and again by almost 5 per cent in 2019. One of the potential explanations for these sudden changes could be conflict between refugees and the local population. Towards the end of 2017, Becharre residents were protesting the increase in the number of Syrian refugees arriving to this District.[29] Following those protests and the murder of a Lebanese woman in a neighbouring municipality by a Syrian,[30] Becharre municipality decided to evict its Syrian refugees from the town.[31] While this event is indicative of events that can explain changes in mobility, further analysis with records from longer and more granular periods of time can allow for more robust testing. The advantage is that CDRs as a data source can fulfil these needs, as data is updated in near real-time with highly granular levels of spatial disaggregation.

Figure 5. Annual mobility in Becharre District



(d) Economic activity

Despite the relatively high levels of mobile phone penetration, reaching about 4.4 million people in Lebanon in 2018[32] (about 64 per cent of the population in the same year),[33] many people still do not have access to a device. In a sense, this is a limitation in the CDR data, as it is unable to capture any information about individuals who do not own cell phones. However, at the aggregate level, not having cell phones can serve as an indicator of underlying socioeconomic conditions. Everything else equal, it can be hypothesised that, for a given geographical area, mobile phone penetration and income display a positive correlation. Additionally, among those who possess a mobile device, the purchasing power of individuals can determine the use they can make of their device. Overall, it is likely that those with a higher income may be able to afford bundles that allow for more calling time. Since calls are being made more and more through apps, the duration of calls and internet use are analysed as proxies for economic activity.

(i) Call duration ratio

The ratio of outgoing to incoming call durations is used to explore users' ability to make longer calls, which can correlate with their economic ability to pay to top up their lines. This ratio is established using the following formula:

$$duration\_ratio = \frac{total\ outgoing\ duration\ per\ district}{total\ incoming\ duration\ per\ district}$$

Figure 6 shows the duration ratio in different districts in the North Governorate in 2016. While the difference between the ratio in different districts is not significant visually, the ratio of outgoing-to-incoming calls is less than 1, meaning that the total outgoing duration is less than the total incoming duration for all North districts.

Figure 6. Ratio of outgoing to incoming duration in the North Governorate (2016)

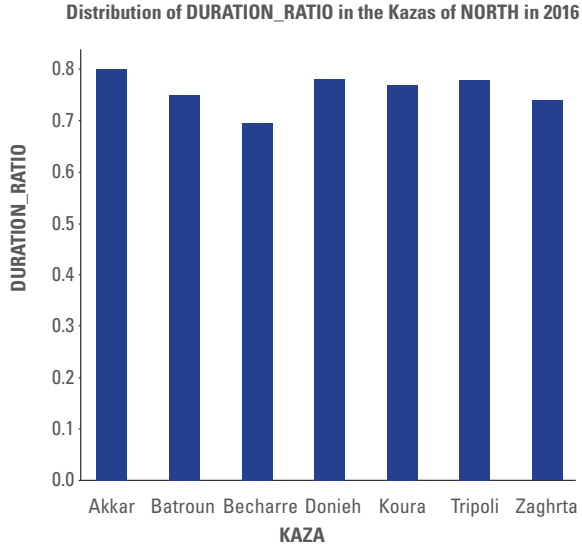
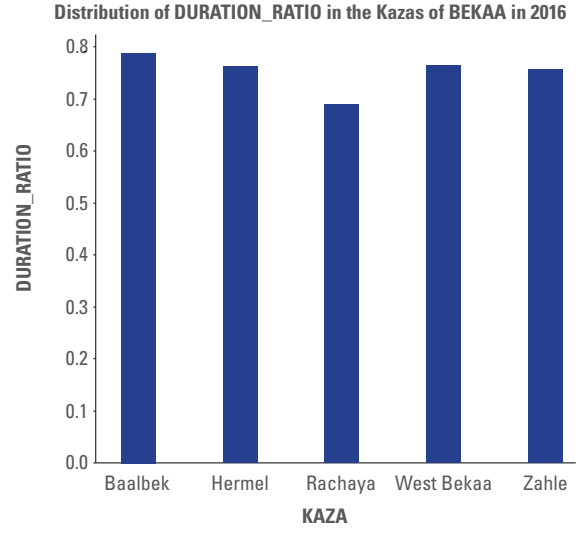


Figure 7. Ratio of the number of outgoing to incoming calls in the Bekaa Governorate (2016)



(ii) Call number ratio

Similarly, the ratio of the number of outgoing to incoming calls is calculated using the following formula:

$$nb.ratio = \frac{total\ number\ of\ outgoing\ calls\ per\ district}{total\ number\ of\ incoming\ calls\ per\ district}$$

Figure 7 shows the call number ratio in districts in the Bekaa Governorate in 2016. The graph shows that the ratio of outgoing to incoming calls is also less than 1, which indicates fewer outgoing than incoming calls, similar to the duration data from figure 6.

(iii) Upload-to-download ratio

In order to reflect online activity, and therefore the capacity to afford it, the upload-to-download ratio is presented. Higher upload means users are actively engaging online, whereas higher download suggests passive activity. The upload-to-download ratio is calculated using the following formula:

$$Up\_Down\_Ratio = \frac{total\ upload\ consumption}{total\ download\ consumption}$$

Figure 8 shows the upload-to-download ratios in the Bekaa Governorate. The graphs show that for all districts, the upload-to-download ratio is very close to 0, which indicates that the upload MB consumption is much lower than that of the download. Here, the key information relies on the proportions, and how these differences can explain the economic conditions in the referenced regions. These results are explored in Section V, where they are contrasted with ground-truth data.

(iv) Refugee daily activity

The data received for Al Tawasol users, proxying Syrian refugee activity, was disaggregated on a daily basis. However, it was not geographically disaggregated. For this reason, the analysis on proxies for refugee activity are made on a temporal basis, with no reference to specific districts. Starting with the hypothesis that differences in calling behaviour can vary depending on working activities, time indicators on weekends, weekdays and holidays are created. In particular, the difference of the daily average call time of Syrian refugees between these days is analysed. Figure 9 shows that holidays have the maximum number of calls, followed by weekdays and weekends, with only slight differences between the three.

Figure 10 displays boxplots of the number of calls for the three types of day. Generally, all three day types are nearly the same, with holidays having higher maximum values (i.e. the most number of calls recorded per day on a given day), and a higher median number of calls for holidays. This suggests that while the amount of calls is elevated on holidays, this difference does not constitute a major deviation from the number of calls from other days. Small differences between weekends and weekdays may be attributed to unemployment, or employment in low-paying jobs.

Figure 8. Upload-to-download ratios in the Bekaa Governorate (2019)

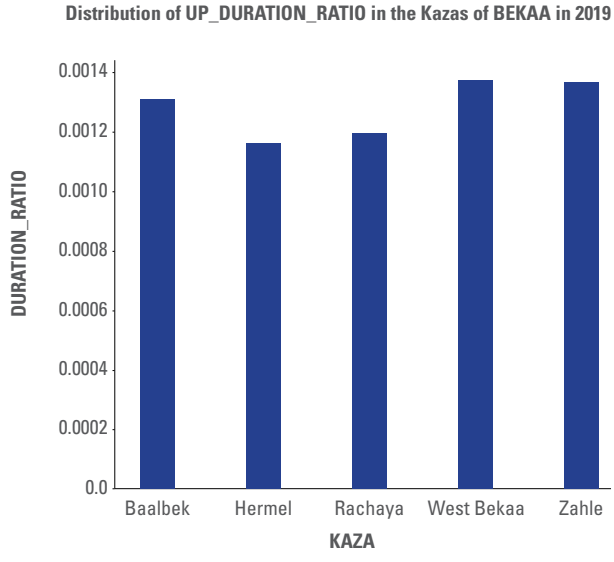


Figure 9. Number of calls by day type

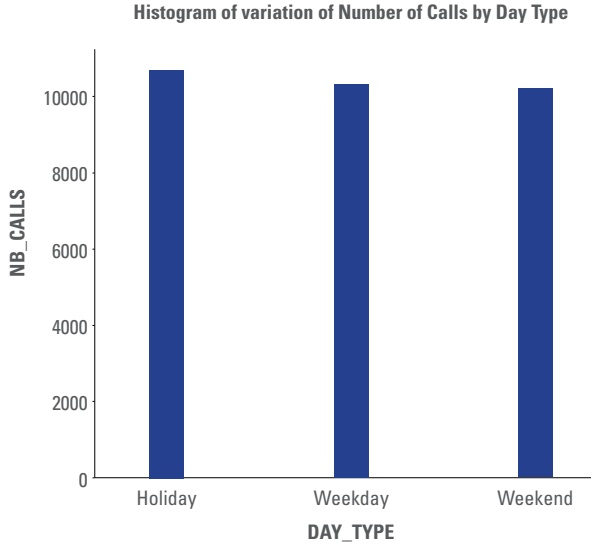
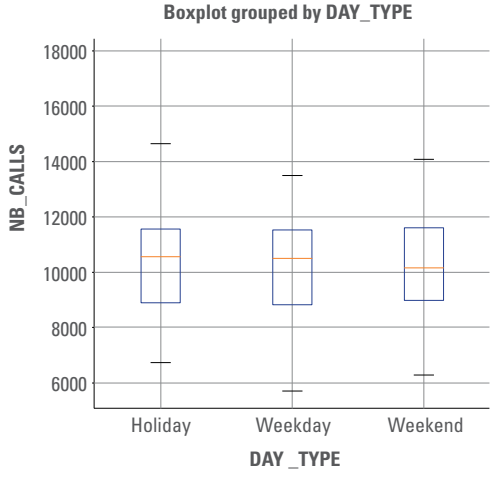


Figure 10. Refugee activity boxplot by day type



(e) Border movement

In this section, we study the variation in total Syrian refugee mobile phone activity. This is achieved by calculating the ratio of the Syrian Touch users (Al Tawasol) to all Touch users using the following formula:

$$refugee.ratio = \frac{total\ calls\ by\ Al\ Tawasol\ users}{total\ calls\ by\ all\ Touch\ users}$$

Figure 11 compares the CDR refugee ratio to the refugee ratio based on official census data, which is calculated based on the number of Syrian refugees as reported by UNHCR in the annual year-end report for 2017,[34] 2018,[35] and 2019[36] and the total Lebanese population.[37] The graph shows that while the two ratios do not exactly match, with the CDR estimates greater than the VASyR's for most periods, both figures display a decreasing trend and overlap in 2019. This could indicate that CDR activity decreased as a result of Syrian refugees leaving the two governorates considered in our analysis. In addition, the VASyR plot points to a change in the rate of decrease in the refugee ratio as of 2018, which is not reflected by the CDR estimate. The proportion of users who moved away from Bekaa and North to other governorates of Lebanon or outside of Lebanon is not known, but it could be better approximated with the inclusion of the other governorates in our analysis.

(f) Security

Population density is used for risk assessment in case of disasters, crises or incidents. This includes planning the support needed in an area based on the population density and determining risk in case of incidents between refugees and the host community.[38] An increase in the population and their activity in a geographical area implies the need to increase the number of mobile sites and cells in that area to accommodate all users in an efficient manner. As suggested by previous work,[39] mobile sites were very densely distributed in urban areas but much less so in rural areas. For this reason, we restrict our attention to the density of sites and cells per district. The distribution of sites and cells in the districts of the North Governorate are plotted in Figure 12.

Potential unexplored methods and uses

Several additional studies have been conducted using CDRs to estimate indicators that have not been covered in this project, mainly due to the nature and the structure of the data on hand. Examples include daytime population estimation, similar to the works undertaken in the Netherlands,[40,41] where the authors studied hourly activity during business hours. Another example is detecting seasonal works, climate-induced migration, etc., as was done in Bangladesh.[42] Other works performed in Estonia[43] and Sri Lanka[44] have looked at commuting statistics, where the researchers proposed relying on locations from cell phone usage to gauge users' homes and work locations to detect commuting flows between the two and their relationship with economic activity and employment.

Figure 11. Percentage of refugees according to CDRs and VASyR

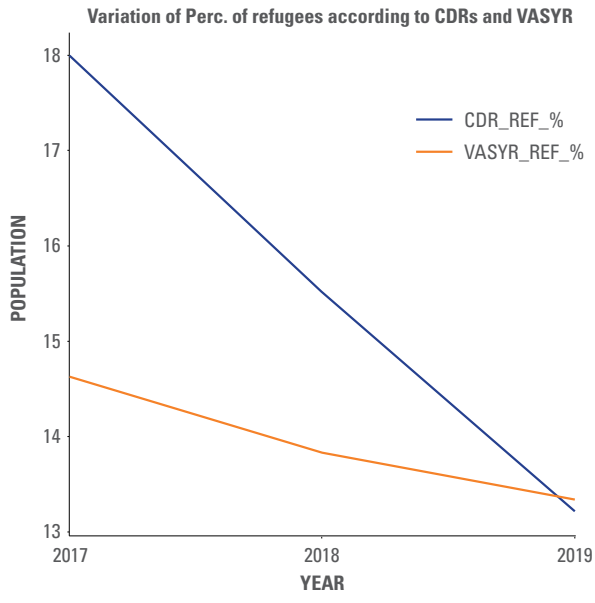
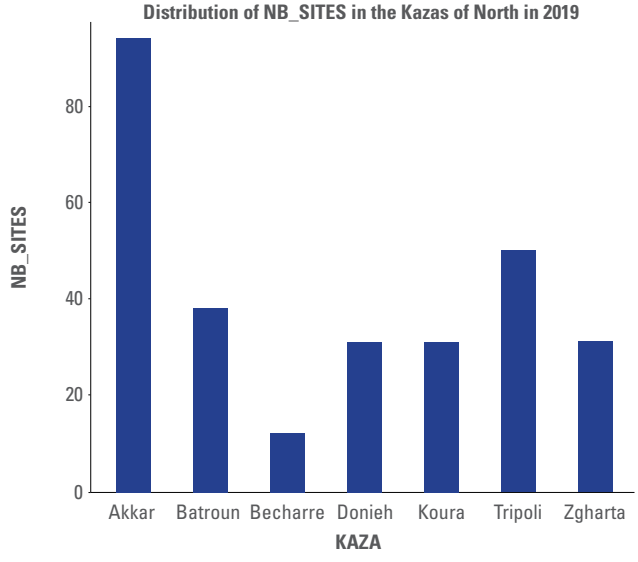
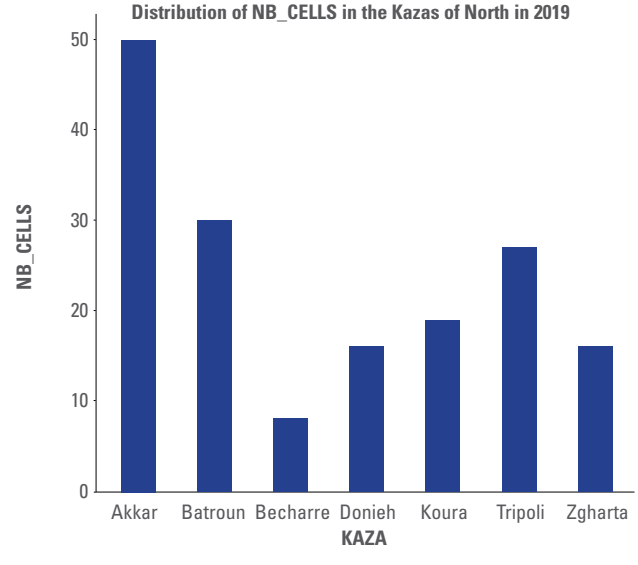


Figure 12. Distribution of numbers of sites for Alfa (top) and cells for Touch (bottom) in the North (2019)





4. Limitations and potential ways to overcome limitations

While CDRs can improve official statistics by increasing timeliness, granularity and accuracy, they cannot replace traditional data collection approaches due to shortcomings inherent in the data source. First, in contrast with official statistics, the data generating process for CDRs cannot be controlled, resulting in non-random sample biases.[45] Second, not every individual is guaranteed to own a single phone only, and thus this mismatch between the unit of observation and the unit of interest may introduce additional bias. Third, not all indicators covered by official statistics are well reflected in mobility or social network behaviour, thus prompting the need for alternative means of data collection. Fourth, for statistical offices to integrate privately held CDRs into statistical production, data access needs to be secured in a sustainable and cost-efficient manner. Fifth, many indicator proxies were established from the obtained CDR data, and these could be better targeted and represented by traditional methods. Specifically, this project faced limitations due to the high level of aggregation applied to the CDRs provided, thereby reducing the richness of terabytes of information to a sample size of 12 observations per year at best, which decreases the statistical power of this study. This fact makes it practically impossible to combine multiple sources of data via multivariate analysis (as even further degrees of freedom will be lost). Going forward, the synergies of CDRs and official statistics should be further exploited through strengthened cooperation and communication between the participating parties in order to. (a) identify specific policy-relevant data challenges that CDRs may help to alleviate (b) improve information retrieval for these analysis tasks while safeguarding privacy. Furthermore, addressing the more source-inherent challenges requires stronger coordination with statistical authorities as many potential biases can be quantified and the sample size can be vastly increased by consensually linking CDRs to survey information at the individual level. Recent cryptographic advancements such as private set intersections[46] present a technological extension to safeguard privacy when combining different data sources.

B. Facebook Ad Platform

1. Description of uses in the literature

In *Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants*,[47] the authors present a proof of concept for the use of Facebook advertising data in estimating stocks of migrants in any country. The authors first find a high correlation between Facebook estimates of the fraction of expats in US state users by country of origin and the latest available data from the American Community Survey. Next, they show a similarly high correlation between expats in the Facebook dataset and the respective foreign-born population from the World Bank (2015) for 96 countries. The authors then discuss the opportunities to use this type of data for demographic research: (i) the data can be obtained via Facebook's Advertising Manager, (ii) it offers an unprecedented depth of dimension, and (iii) it includes multiple details about the population, ranging from demographics to socioeconomic and educational data, all of which are regularly updated. The authors suggest that the dataset can be perceived as a continually updated census. Subsequent work has similarly highlighted the value of this data to obtain up-to-date estimates of stocks of migrants in different host countries.[48]

2. Use for this particular setting

In our approach, we used Facebook's advertising platform figures on monthly active Facebook users to investigate the economic, demographic and educational conditions of Syrian refugees and host communities.

We collected estimates of Monthly Active Users (MAU) of Facebook for a variety of attributes and their combinations,

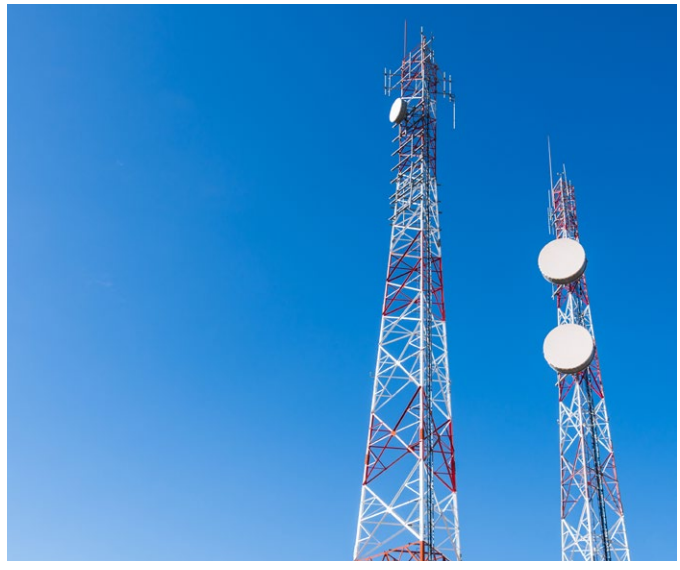
at various geographic resolutions in Lebanon (governorate, district and city levels). The attributes for which data was collected include the following:

1. Age: age group of the users.
2. Gender: male, female or all.
3. Language: language in which users use Facebook (Arabic, English, all).
4. Lives abroad (expat status): everyone (expat or non-expat), all expats, all expats but not from specified Arab countries (these countries are: Morocco, UAE, Saudi Arabia, Jordan, Lebanon, Algeria, Kuwait, Qatar; while this is not a complete list of Arabic speaking countries, these are the only countries we could explicitly exclude with the Facebook API) and non-expats.

5. Education: self-declared educational attainment of users.

6. Device/Network types: the device and network types with which users access Facebook. These include:
a. Network access: 2G, 3G, 4G or WiFi;
b. Mobile OS: iOS, Android, Windows or other;
c. High-end phones: latest Samsung or iPhones;
d. Device types: A variety of other device types such as tablets, smartphones, Cherry Mobile etc.

Language and expat status are used to define the host and refugee communities. Since Facebook API does not allow us to specifically specify expats from Syria, we define refugee communities as the Facebook users who use Facebook in Arabic and are expats but not from the above specified Arabic countries. A similar approach can be found in an earlier study examining the assimilation of Syrian refugees in Germany.[49] The non-expat users represent the host community.



3. Results

(a) Population pyramid

Figure 13 shows the age and gender distribution of Facebook users aged over 13 years old for the host and refugee communities respectively.

We observe a similar distribution shape per age for all categories, with the distribution being right skewed and peaking at the age range of 25 to 29, followed by ages 20 to 24. This finding is not surprising as individuals falling into this category are expected to constitute the largest group of Facebook users. Contrasting both communities' histograms for either gender, we find than an even larger proportion of users of young age exists in the refugee community (for instance 71 per cent of males in the host community are under 40 as opposed to 86 per cent in the refugee community), which again is to be expected.

Figure 13. Facebook population pyramid for host (left) and refugee (right) communities

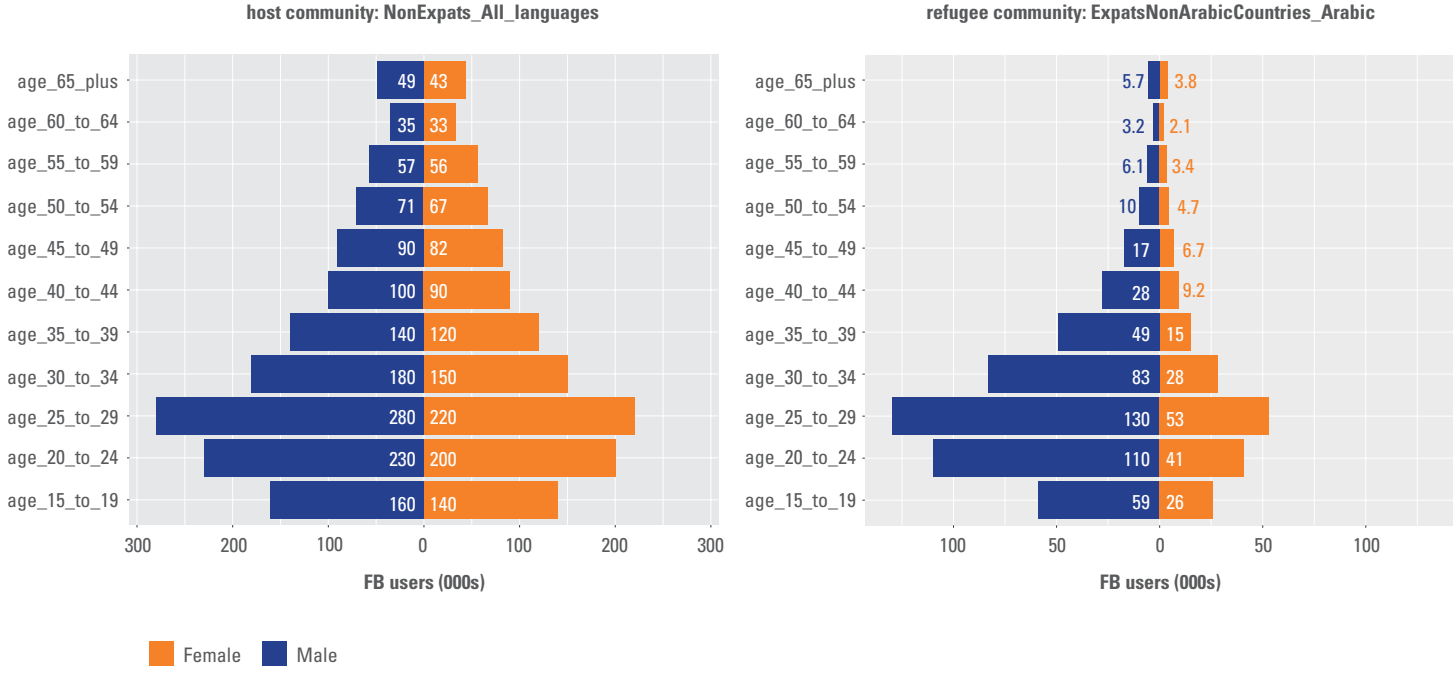


Table 1. Target and non-traditional variables explored in this work

Expat_status	language	age_13_plus_all_genders
All	All_languages	3,500,000
NonExpats (host communities)	All_languages	2,600,000
ExpatsNonArabicCountries (Syrian Refugees)	Arabic	720,000

(b) Refugee distribution

The table below summarises the Facebook Monthly Active Users for the different types of users.

(c) Economic activities

Figure 14 shows the distribution of the type of network access of Facebook users for host and refugee communities, which reveals that over 50 per cent of the network types used are other than 2G, 3G and 4G Networks or Wifi. Note that some categories may overlap (e.g. some users may use both 3G and 4G network).

Figure 15 shows the distribution of the type of Mobile OS of Facebook users for host and refugee communities, where again we notice that over 50 per cent of mobile OS types are none of iOS, Android or Windows.

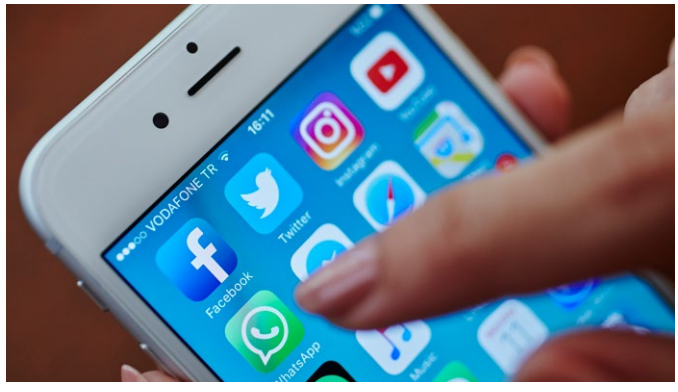


Figure 14. Network type of Facebook users

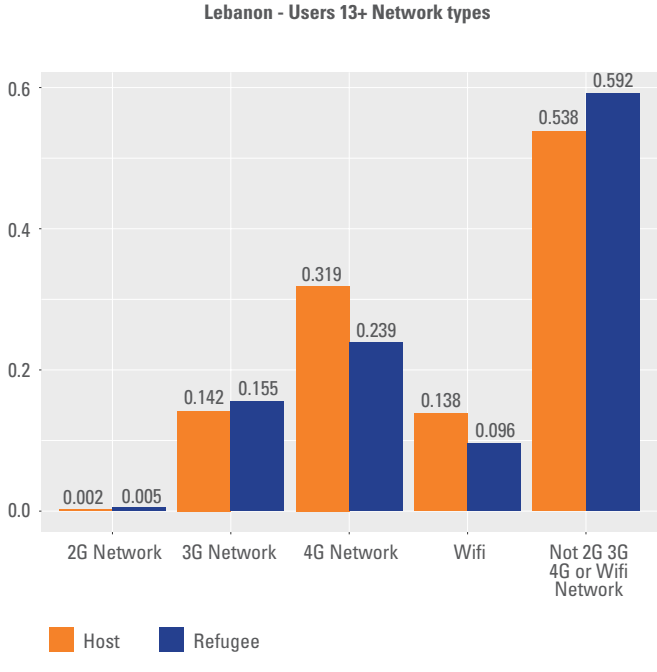


Figure 15. Mobile OS type of Facebook users

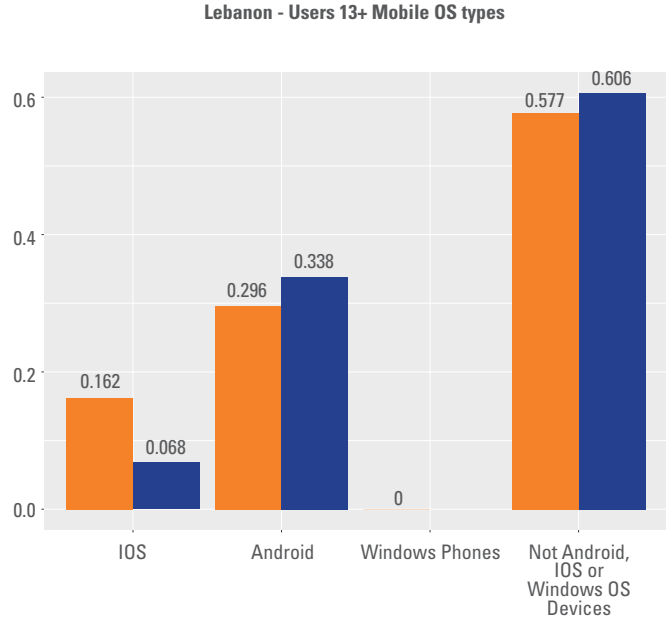


Figure 16 shows the distribution of the device type of Facebook users in the host and refugee communities.

(d) Education

Figure 17 shows the distribution of the self-declared education of Facebook users in the host and refugee communities. As we can see from the graph, about 60 per cent of Facebook users do not declare their education. In general, the figures for the self-declared education are similar for each category between both communities, but the refugee community seems to have a slightly lower fraction of users with higher levels of education.

Figure 16. Mobile devices of Facebook users

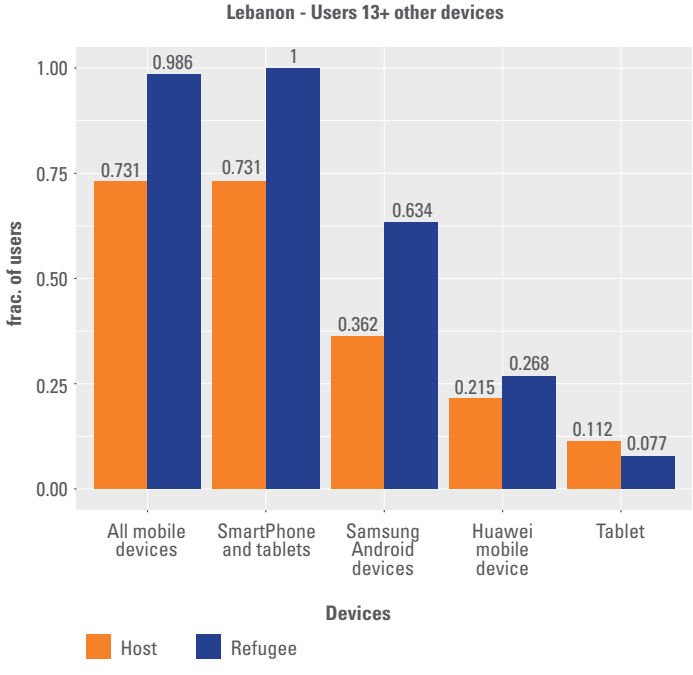
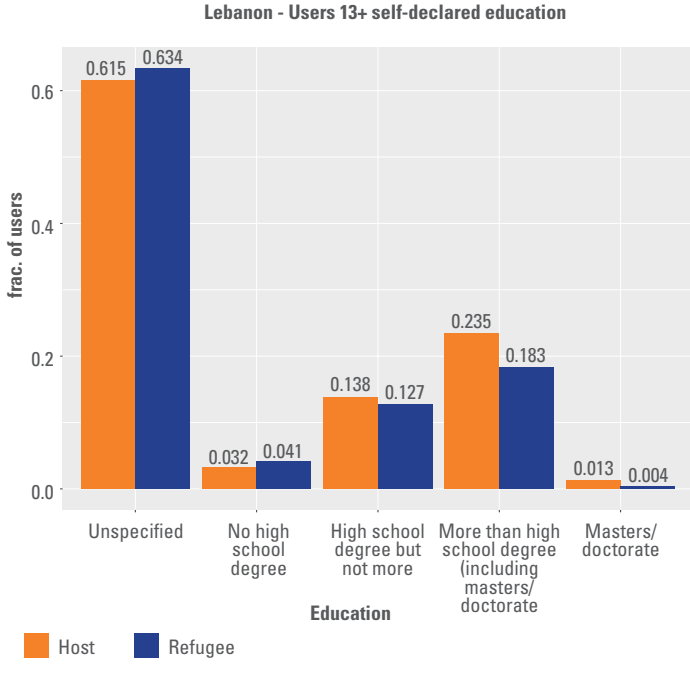


Figure 17. Self-declared education of Facebook users



4. Limitations and potential ways to overcome limitations

The main limitation of this approach is that individuals who are expats from Syria cannot be targeted directly on Facebook, which led us to use Arabic speaking expats who were not from other Arabic speaking countries which the Facebook platform supports as a proxy for Syrian refugees on Facebook. This leaves some uncertainty as to how other populations such as Palestinian refugee populations may be appearing in these data.

Another important consideration is data sparsity, especially when collecting data at finer spatial resolutions. While the FB platform allows a broad range of targeting options including fine-grained location targeting, we encountered sparsity problems when collecting data for the finer resolution city level: out of 770 cities, 157 (20 per cent) had a host population of users greater than 1000 and 49 (6.3 per cent) had a refugee population greater than 1000.

In terms of gender distribution, a larger gender imbalance was observed for the Syrian refugee population on Facebook in Figure 13 with 2 to 3 times more men than women. In addition, our analysis does not reflect qualitative offline behaviour such as potential sharing of phones/social media accounts in a household for example. According to the 2018 VASyR report Facebook is used by about 16 per cent of Syrian refugee households: "More than three quarters (79 per cent) of refugee households were active on social media. The most utilised digital platform by far was WhatsApp (78 per cent), trailed by Facebook (16 per cent)".

C. Global Database of Events, Language, and Tone (GDELT)

1. Description of uses in the literature

In this approach, we study the sentiment of the media towards Syrian refugees. This varies from articles reporting on incidents, humanitarian aid, employment and economic issues to other kinds of events involving Syrian refugees.

Several studies have been performed on using media and social media content and sentiment for predicting social unrest. *In Predicting Social Unrest Using GDELT* [50], the authors look at the extent to which news media data can detect social unrest. They extracted articles from GDELT for the USA based on a set of themes in addition to extracting major social unrest events for the same period. According to the authors, huge social unrest events might be preceded by numerous reports of relevant themes in the media. The authors then train a machine learning model on the GDELT articles to predict the matched social unrest events. Using the trained model and newer unmatched GDELT articles, they managed to predict locations that would be subject to social unrest as of the date of the writing of the paper.

Based on the similar assumption that the eruption of social unrest is preceded by a sequence of events, the authors of *Predicting Social Unrest Events with Hidden Markov Models Using GDELT* [51] propose a machine learning model to predict significant social unrest events on a country level based on articles they extracted from GDELT.

In *Integration of Syrian Refugees: Insights from D4R, Media Events and Housing Market Data* [52], the authors combine CDR data with geo-localised events related to refugees extracted from GDELT. This is done using Turkish CDRs and extracting province-level events from GDELT. The results show that for province-level effects, refugee-related events are correlated with increased native call volume, rather than increased refugee call volume.

2. Use for this particular setting

As stated earlier, GDELT extracts details about each article collected in terms of events and sentiment. Since we are studying the relation between Syrian refugees and host communities in Lebanon, we are interested in the actors, location, article URL and year of the events. Table 3 describes the details we extracted from GDELT.

(a) Extracting articles of interest

We extracted articles about events involving Syrians and taking place in Lebanon by specifying that one of the involved actors' country is Syria. Since we are interested in the relations between Syrian refugees and Lebanese host communities, we made sure that the extracted articles involve Lebanese actors as well. This is achieved by extracting articles where either Actor1 is Lebanese and Actor2 is Syrian, or vice versa, where the event's country is Lebanon.

We use Google Cloud Platform's BigQuery, which is the serverless data warehouse where GDELT is stored, to extract the articles of interest from GDELT. Google BigQuery allows the storage of huge databases and

Table 3. Event attributes extracted from GDELT

Column	Description
Year	The year the article was published
Actor1CountryCode	3-character code for the country of Actor1
Actor2CountryCode	3-character code for the country of Actor2
ActionGeoFullName	Full human-readable name of the matched location. It is in the format of City/Landmark, State, Country
ActionGeoCountryCode	2-character country code for the location
ActionGeo_Lat	Centroid latitude of the landmark of ActionGeoFullName
ActionGeo_Long	Centroid longitude of the landmark of ActionGeoFullName
SourceURL	URL of the article
lang	Language of the article (not available for Arabic)
score	Sentiment score ranges between -1.0 (negative) and 1.0 (positive) and corresponds to the overall emotional leaning of the text (not available for Arabic)

enables the querying of these databases using SQL statements. For each of the years 2016-2019, we query GDELT's events and sentiment databases to extract the articles of interest, along with their links, exact location, and sentiment. This is achieved using the following SQL statement:

```
SELECT DISTINCT(sourceurl),ActionGeoFullName,ActionGeo_Lat,ActionGeo_Long,lang,score
FROM gdelt-bq.gdeltv2.events JOIN gdelt-bq.gdeltv2.geg_gcnlapisent
ON gdelt-bq.gdeltv2.events.SOURCEURL = gdelt-bq.gdeltv2.geg_gcnlapisent.URL
WHERE YEAR = 'YEAR' AND ActionGeo_CountryCode = 'LE'
```

Since GDELT stores events and sentiment in different databases while storing the link of the articles with each row in both databases, we join the databases on the common column, the source URL, to end up with the sentiment and the event details in one final resulting table. We then export the table into a CSV file for further processing and analysis.

(b) Filtering the articles

Since we extracted articles that involve Syrian actors, there might be a few articles reporting on Syrian actors who are not necessarily refugees. In addition, for those articles reporting on Syrian refugees, we would like to categorise them to understand in what context Syrian refugees are being mentioned.

For this reason, we created a scraping tool that automates extracting the content of the articles. This tool works by visiting the source URL that was extracted from GDELT. Based on a set of keywords tied to Syrian refugees, the tool only keeps articles that mention Syrian refugees.

As mentioned earlier, the Google Cloud NLP Sentiment API used by GDELT to extract articles' sentiment and language did not support Arabic as of the time of writing this policy note. For this reason, we also needed to identify which articles were in Arabic. For English articles, we already have the language extracted from GDELT. For this reason, the scraping tool also extracts Arabic content from the pages whose language was not identified to be English, and if no Arabic content is found the article is dropped. Otherwise, the article's language is set to Arabic.

In addition to verifying that the article mentions Syrian refugees, the scraping tool checks if the article is reporting on various categories. This is also achieved by defining a set of keywords for each topic and checking whether the article contains any of these keywords. Table 4 summarises the topics and their respective keywords.

These keywords were also translated to Arabic in order to match Arabic articles. Note that one article may contain keywords from different topics and can therefore belong to more than one category.

(c) Arabic sentiment analysis

Since GDELT does not provide sentiment analysis for Arabic articles, we use an open source Arabic sentiment analysis API called Mazajak[53] to categorise the sentiment of the Arabic articles. Mazajak labels a given text as one of three options: positive, negative or neutral.

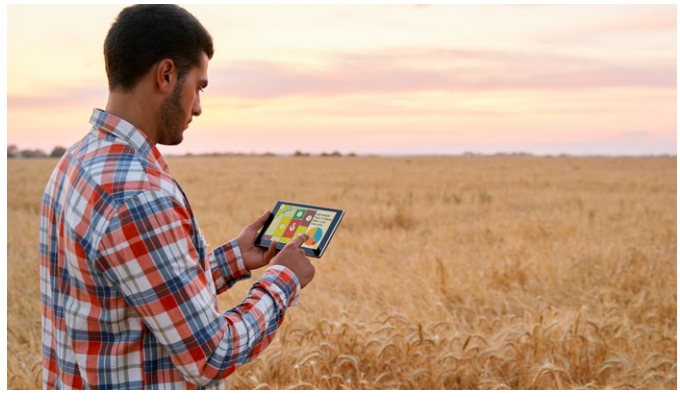


Table 4. Scraped categories, topics and keywords

Category	Topic	Example Keywords
Economic Activity	Employment	Unemployment, labour force, labour market, hiring, firing, work
	Poverty	Poverty, scarce, hunger, displacement
	humanitarian aid	Humanitarian, aid, assistance, cash transfers, vouchers, e-cards, support
Security	Incidents	Discontent, uprise, protest, violence, attack
	Shelters	Camps, shelters, displacement, expulsion
	Demands of return of Syrian refugees to Syria	Return to Syria

(d) Extracting district from article location

GDELT's geotagging of articles is as detailed as the landmark where the event took place, along with the latitude and longitude. In order to perform a regression as we will see in the results section, we converted the detailed geotagging to district-level tagging, by matching each GDELT action location's latitude and longitude with the closest district's latitude and longitude.

3. Results

In this section, we look at the variation in Arabic and English articles reporting on the topics of interest across the years, along with the variation in the sentiment of these articles. For Arabic articles, sentiment is reported as a percentage distribution of the three labels: positive, negative and neutral. For English articles, sentiment is reported as a score between -1 (negative) and +1 (positive). Our findings are summarised as follows:

1) Variation in the number of articles per topic:

a) Across all four years studied, 2017 witnessed the most coverage dedicated to Syrian refugees on all topics, except for poverty. This could be triggered by political events concerning the refugees, which in turn influenced their socioeconomic status and thus sparked more notice from the media.

b) An increase is observed in the number of articles in English over the years regarding Syrian refugees, while a decrease occurred in the number of articles in Arabic. This could signal that more attention is being paid to the conditions of refugees abroad than in Arab countries.

2) Sentiment of articles per topic:

a) An increase in negative sentiment on all topics, and hence an increase in overall negative sentiment, is seen in our results.

b) An increase in positive sentiment on all topics is found, but this is offset by an equal or larger increase in negative sentiment.

c) A decrease in neutral sentiment is found, which implies that the changes in negative sentiment dominates opposing changes in the other two sentiments. The decrease in neutral sentiment could indicate for instance that authors are becoming more polarised over the years with regards to Syrian refugees.

(a) Refugees

We retrieved a total of 7400 unique articles reporting on Syrian refugees in the years 2016-2019.

Figure 18. Number of articles on Syrian refugees, 2016-2019

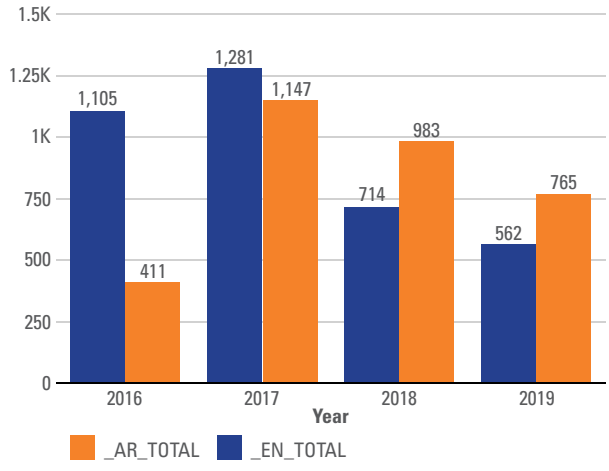
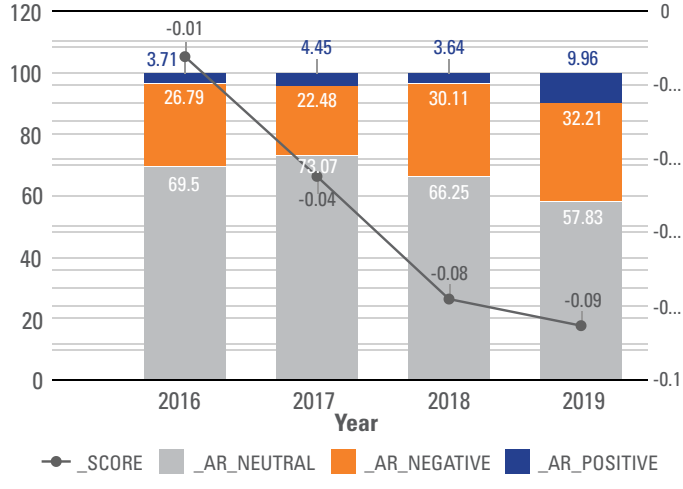


Figure 19. Sentiment of articles on Syrian refugees, 2016-2019



(b) Economic activities

(i) Employment

Figure 20. Number of articles on Syrian refugees and employment, 2016-2019

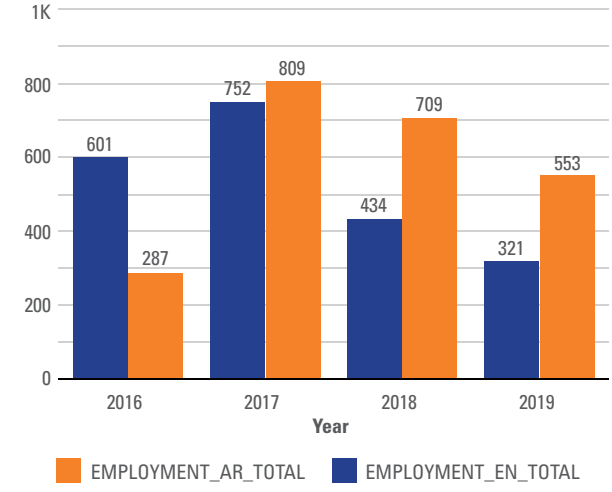
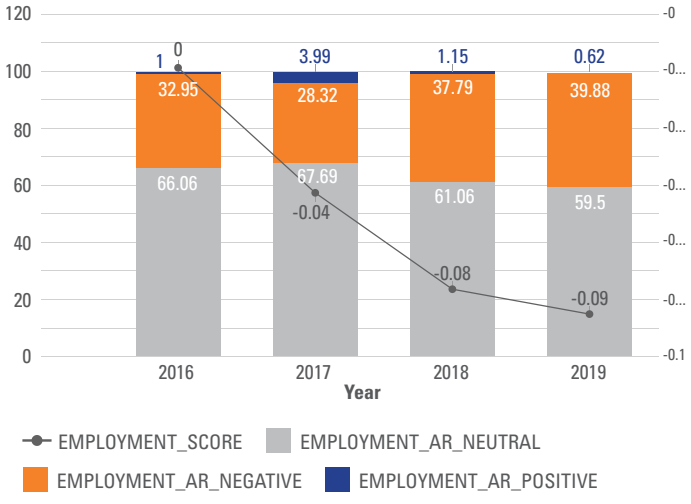


Figure 21. Sentiment of articles on Syrian refugees and employment, 2016-2019



(ii) Poverty

Figure 22. Number of articles on Syrian refugees and poverty, 2016-2019

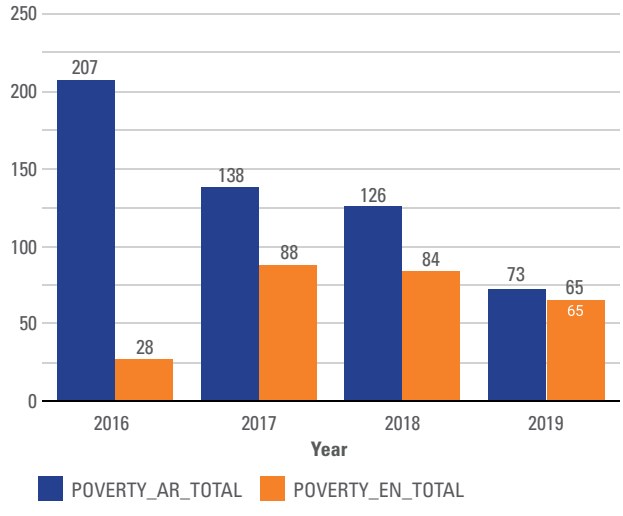
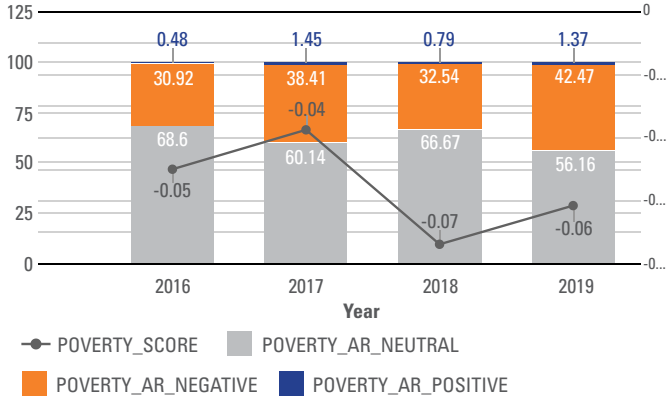


Figure 23. Sentiment of articles on Syrian refugees and poverty, 2016-2019



(iii) Humanitarian aid

Figure 24. Number of articles on Syrian refugees and humanitarian aid, 2016-2019

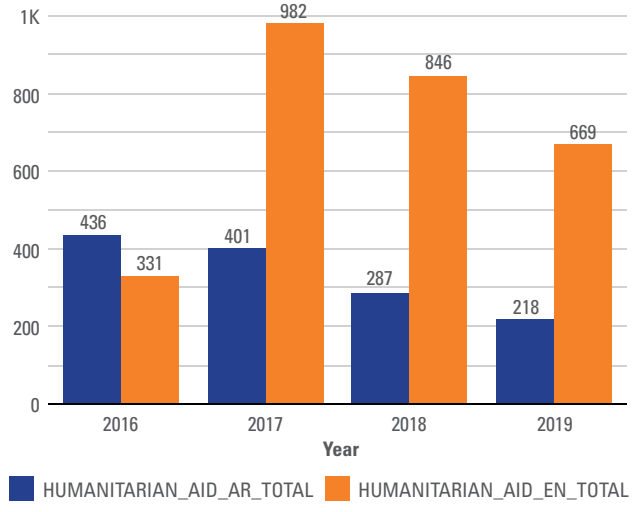
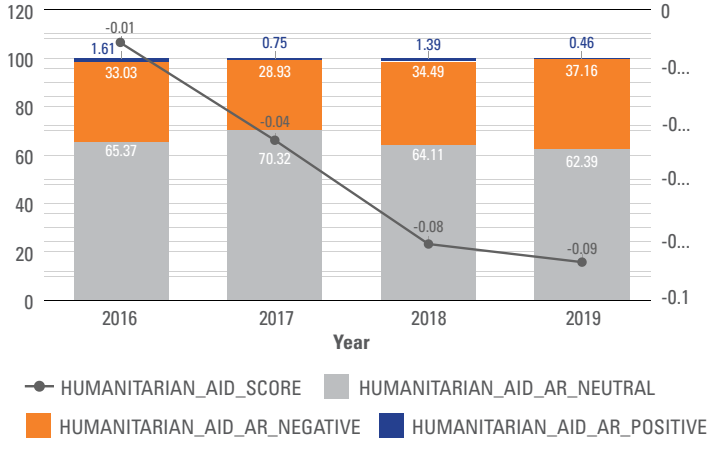


Figure 25. Sentiment of articles on Syrian refugees and humanitarian aid, 2016-2019



(b) Security

(i) Incidents

Figure 26. Number of articles on Syrian refugees and incidents, 2016-2019

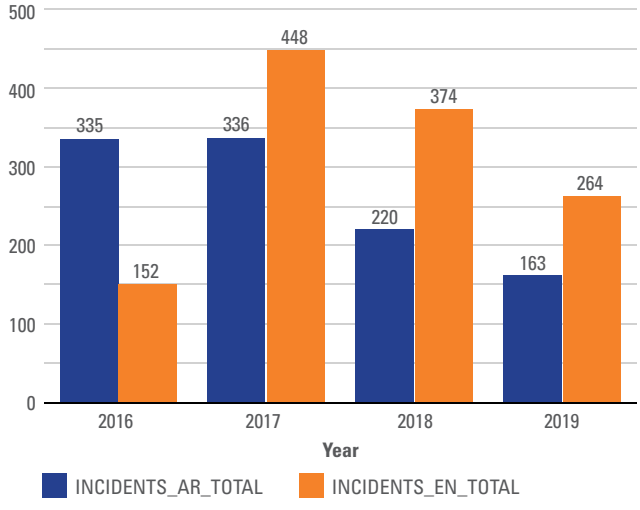
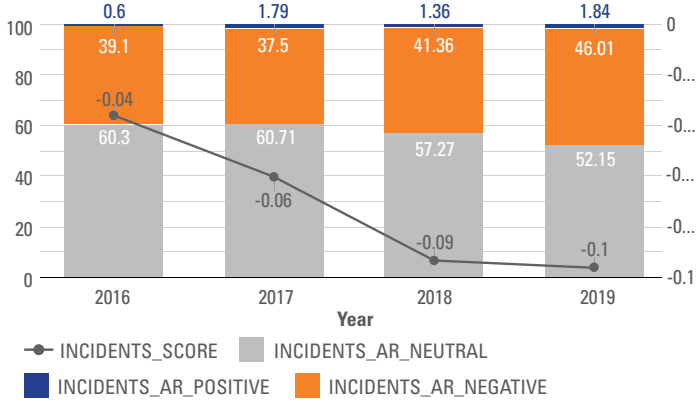


Figure 27. Sentiment of articles on Syrian refugees and incidents, 2016-2019



(ii) Shelter

Figure 28. Number of articles on Syrian refugees and shelter, 2016-2019

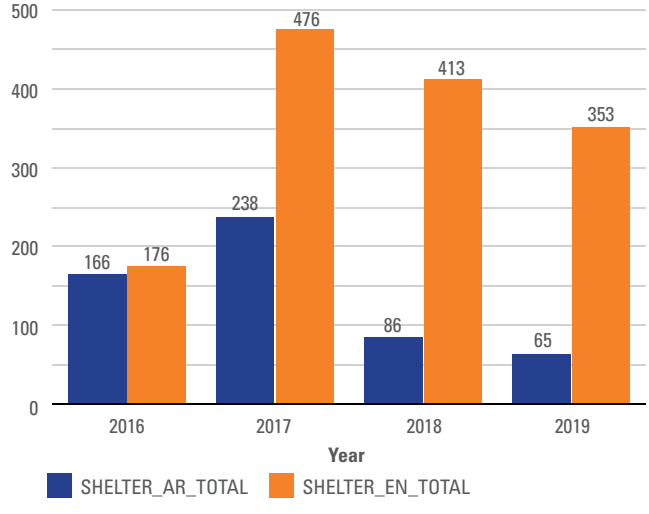
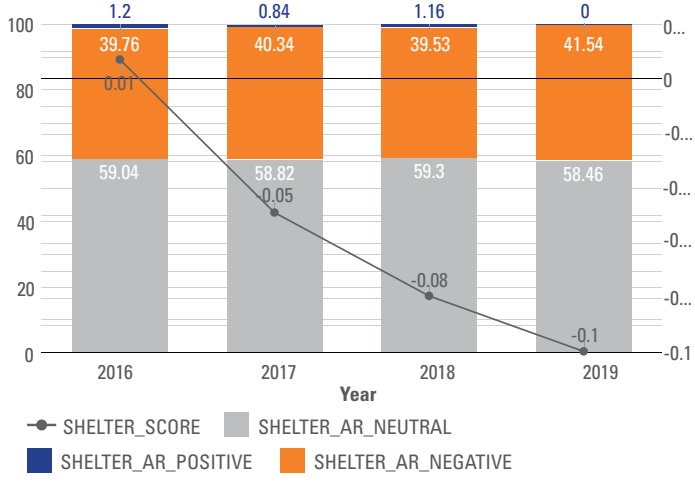


Figure 29. Sentiment of articles on Syrian refugees and shelter, 2016-2019



(iii) Demands of return of Syrian refugees

Figure 30. Number of articles on Syrian refugees and demands of return, 2016-2019

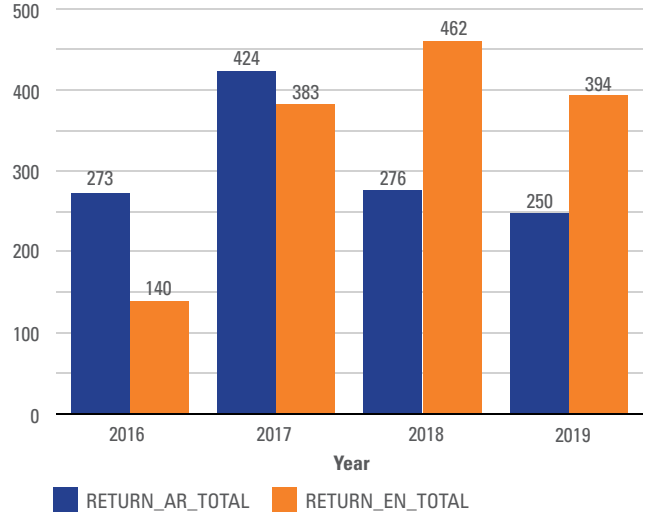
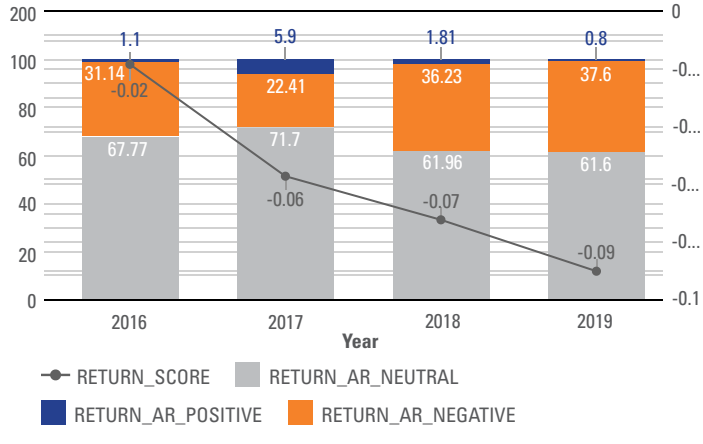


Figure 31. Sentiment of articles on Syrian refugees and demands of return, 2016-2019



4. Potential unexplored methods and uses

A great deal of text-based research and media analysis can be performed using GDELT. These include crowdsourcing extracting methods, where online workers are asked via crowdsourcing platforms such as *Amazon Mechanical Turk* [54] to manually extract pieces of information from each article, whether it involves performing a simple copy and paste from the article or selecting options based on the article text, to name a few. These manually extracted pieces of information can then be used to train machine learning models to automate the extraction process.[55] However, these methods were not relevant to the scope of our work, so they were not explored.

Limitations and potential ways to overcome limitations

The main limitation of using media sentiment analysis and frequency is that it subjects our work to the bias of news outlets. This can happen due to certain sources over-reporting on specific topics, whether positively or negatively, and therefore affecting our findings. One potential way to deal with this is to narrow the number of articles retrieved from each source. This limit on article retrieval can be topic-based or time-interval based, to prevent one source or another from affecting our findings, therefore ensuring equal coverage from all news sources for each topic studied.

Another main limitation is the fact that media can be controlled by political parties, governments, organizations and others, which might exert a bias on the dataset collected from the web-scraping tool and thus skew the obtained results. A good way to avoid this is to integrate social media into this kind of work to get the unfiltered opinion of the people, rather than the media.

C. Twitter

1. Description of uses in the literature

In an approach similar to our GDELT approach, we study the sentiment of Twitter users towards Syrian refugees.

In *Analysing Refugee-Host Community Narratives on Social Media*,[56] the authors used Twitter APIs and a set of keywords and phrases to retrieve Arabic and English tweets posted in a single month in Lebanon about Syrian refugees. From these tweets, the authors extracted sentiment, events and topics and analyse the trends in sentiment in the dataset based on the events. Similarly, in *Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis*,[57] the authors extracted Turkish and English tweets from Twitter, performed a sentiment analysis in both languages, and calculated the frequencies of terms pertaining to Syrian refugees for both languages using word clouds.

In *Social Media and Forced Displacement: Big Data Analytics & Machine-Learning*,[58] the authors proposed a near-real time monitoring system that monitors Twitter for tweets about refugees in Europe. They trained machine learning models to classify tweets and extract their sentiment to be used in the real-time monitor. The authors suggest that understanding the sentiment of the public in a near-real time monitoring system can help inform operational responses in support of the Europe emergency regional protection strategy.

2. Use for this particular setting

In this work, we intended to study Twitter in the same manner as we did for GDELT. For this reason, we made use of the Twitter API's Tweet searching feature, specifically the Search Tweets API: Full Archive.

Twitter's Full Archive API allows developers to request tweets dating back to 2006 using customised search queries. Each query returns up to 500 tweets. In the queries, developers can specify keywords, languages, the tweeter's country, the date range to be searched and more. The following are the operators that we are interested in in our Twitter approach:

Table 4. Scraped categories, topics and keywords

Operator	Description
Keyword	Matches a tokenized keyword within the body or URLs of a Tweet
Lang	Language of tweet. English and Arabic supported
Profile_ country	Country of the user who posted the Tweet as specified in his/her profile
Fromdate	Start date to be scraped
Todate	End date to be scraped

Method

Scraping twitter

Based on the above, we build search queries that scrape Twitter for tweets about Syrian refugees in the context of the same topics and keywords as we did for GDELT. Since Twitter returns up to 500 tweets per query, we queried each of the 7 topics twice, once for English and once for Arabic, in addition to two queries about Syrian refugees in general. The following is an example query:

(syrian OR refugee) (incident OR discontent OR violence OR assault OR explosion)

profile_country:LB -is:retweet lang:en

The above query looks for tweets that mention Syrian refugees and incidents, discontent, etc… which are the keywords we chose for the security topic, as mentioned in the GDELT section. The query also specifies that the tweet's poster's current country has to be Lebanon, and the tweet must be in English. We also add the -is:retweet operator, which ensures that we do not fetch retweets and therefore not include a single tweet more than once in our dataset.

The above query structure was sent as a request to Twitter's Search API full archive, along with a fromDate of 1 January and toDate of 31 December for each of the years 2016-2019. The request is repeated for each year, for Arabic and for English tweets, for each of the topics of interest.

Sentiment analysis

For Arabic sentiment analysis, we use the same open source Arabic sentiment analysis API that we used for GDELT (Mazajak) to measure the sentiment of the Arabic articles. For English sentiment analysis, we use Google Cloud NLP Sentiment API, which is used by GDELT to record the sentiment of articles. Therefore, similar to GDELT, the sentiment of Arabic tweets will be either "positive", "negative" or "neutral", while the sentiment of English tweets will be a score that ranges between -1 (negative) and +1 (positive).

3. Results

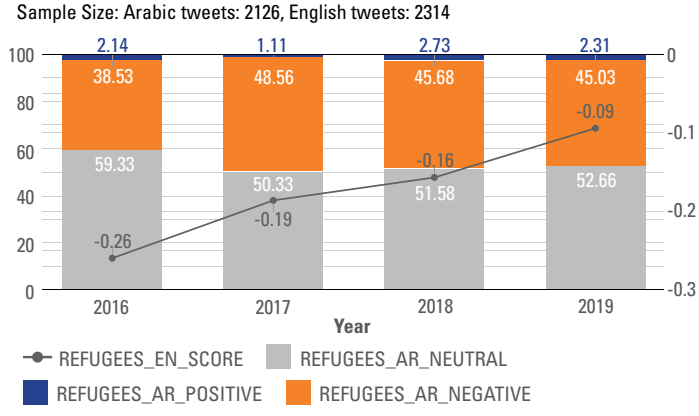
In this section, we look at samples of Arabic and English tweets mentioning the topics of interest over the years, along with the change in the average sentiment of these tweets. Similar to the GDELT results, the sentiment of Arabic tweets is reported as a percentage distribution of positive, negative and neutral, while the sentiment of English tweets is reported as a score between -1 (negative) and +1 (positive).

(a) Refugees

We retrieved a total of 4440 tweets mentioning Syrian refugees in the years 2016-2019.

An increase is observed in the number of articles in English over the years regarding Syrian refugees, while a decrease occurred in Arabic counterpart. An increase in negative sentiment on all topics, from retrieved 7400 unique articles in the years 2019-2016.

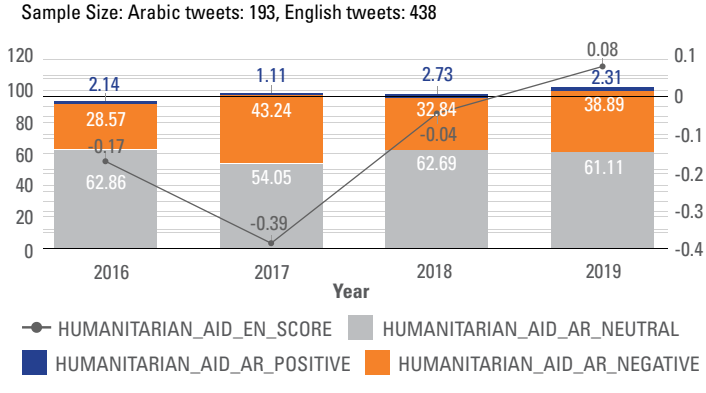
Figure 32. Sentiment of tweets mentioning Syrian refugees, 2016-2019



(b) Economic activities

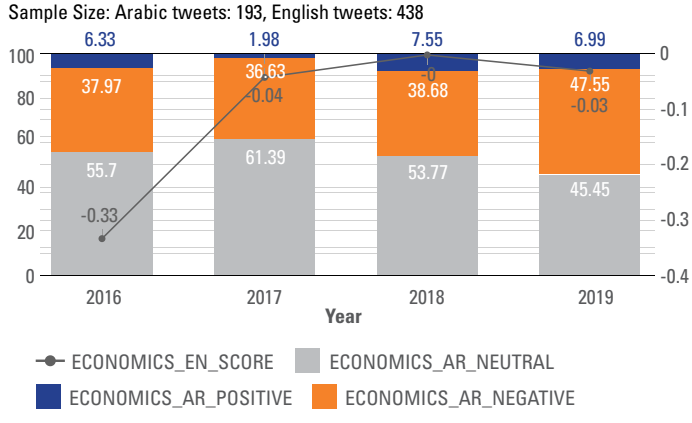
Humanitarian aid

Figure 33. Sentiment of tweets mentioning Syrian refugees and humanitarian aid, 2016-2019



a. Employment

Figure 34. Sentiment of tweets mentioning Syrian refugees and employment, 2016-2019



b. Poverty

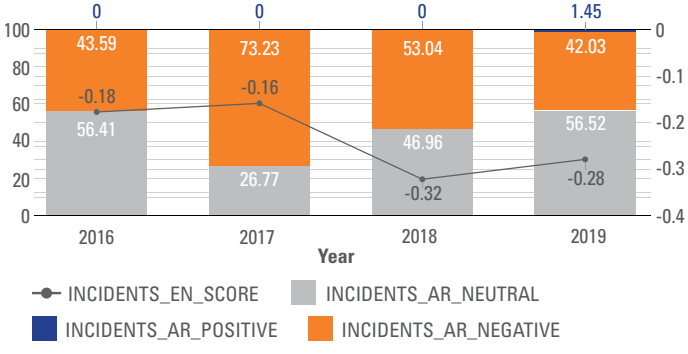
The tweets retrieved about Syrian refugees and poverty had a very small sample size of only 29 Arabic tweets and 254 English tweets, so we dropped them from the analysis.

(c) Security

(i) Incidents

Figure 35. Sentiment of tweets mentioning Syrian refugees and incidents, 2016-2019

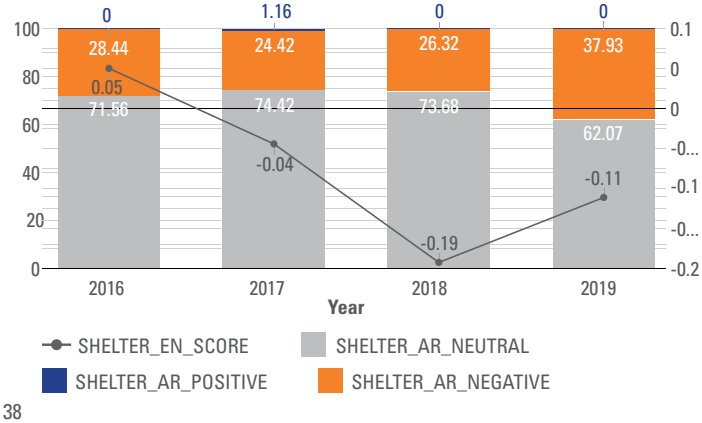
Sample Size: Arabic tweets: 389, English tweets: 316



(ii) Shelter

Figure 36. Sentiment of tweets mentioning Syrian refugees and shelter, 2016-2019

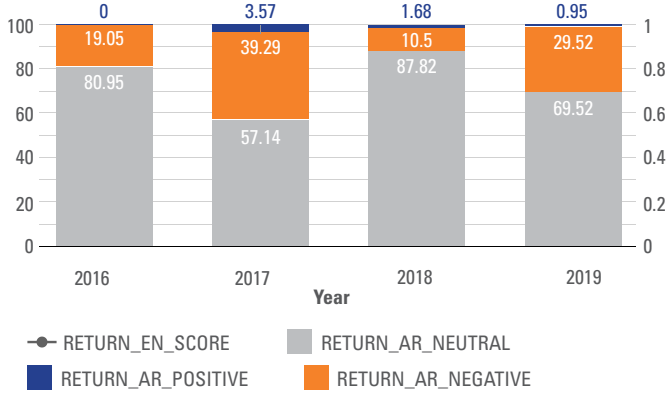
Sample Size: Arabic tweets: 310, English tweets: 402



(iii) Demands of Return of Syrian refugees

Figure 37. Sentiment of tweets mentioning Syrian refugees and demands of return, 2016-2019

Sample Size: Arabic tweets: 420, English tweets: 325



4. Potential unexplored methods and uses

Similar to GDELT, a lot of text-based research and social media analysis can be performed using Twitter. This can also include extracting pieces of information from tweets.[59] However, also similar to GDELT, these methods were not relevant to our scope of work. They also rely on information extracted from people tweeting, which could be subject to bias, bots, etc. and oblige us to make too many assumptions. Therefore, these methods were not explored in this work.

5. Limitations and potential ways to overcome limitations

Unlike GDELT, scraping Twitter allows us to get the raw and unfiltered opinions of the people tweeting about topics, thus decreasing the risk of media bias. However, the main limitation of using social media sentiment analysis and frequency is that it subjects our work to tweets published by bots, spammers, etc. and thus the risks of fake and/or biased data. Similarly, to the GDELT solution, one potential way to deal with this is to limit the number of tweets

retrieved per user. The Twitter API responds with the user public metadata as well as the tweet content, allowing us to filter tweets per unique user. However, this results in more requests to the API and fewer tweets in our sample.

4

Predictive Capacity of Data Sources with Ground Truth Data

A. Linear univariate and multivariate regression

In this section, we look at the results of the linear regression that we performed for each ground-truth label and its relevant indicators calculated in our approaches.

1. Population and refugee distribution

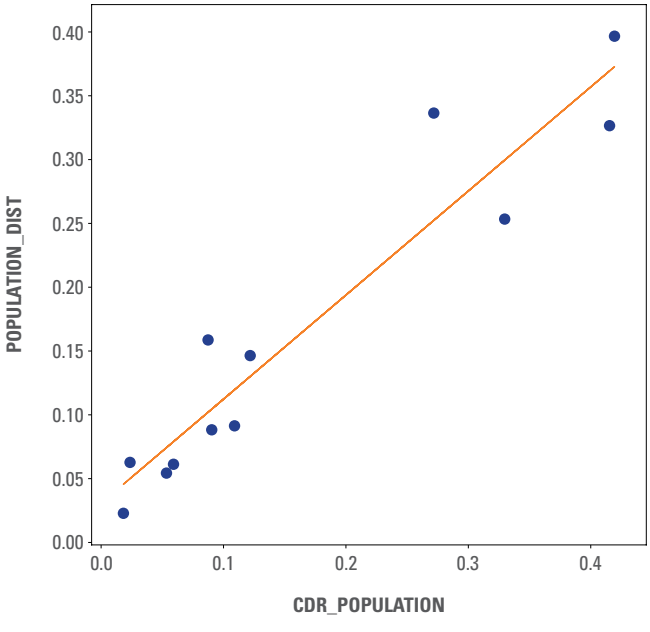
(a) CDRs

We performed a linear regression between the CDR population indicator from Alfa and Touch separately and the population distribution of the official statistics of CASLFS. Since our population proportion is calculated as the per cent of each governorate in each district, we convert the population proportion provided by CASLFS into a similar structure using the following formula:

$$district.population.prop = \frac{district.population}{governorate.population} \times 1000$$

Figure 38 shows the linear regression[60] of the CDR population indicator and the CASLFS population.

Figure 38. Linear regression of population distribution of the 12 Kazas



R^2 score: 0.9

Mean Squared Error: 0.002

Table 6 shows the correlations between the population indicator as calculated using Alfa and Touch CDRs and the population as per CASLFS. The table also shows a high correlation between the number of sites and cells (Alfa sites and Touch cells) versus the population proportion. Therefore, the higher the sites and cells in an area, the higher the population.

Table 6. Pearson correlation between CDR indicators and CASLFS population distribution

	Pearson correlation
Alfa Population Indicator	0.95
Touch Population Indicator	0.88
Touch Number of Cells	0.89
Alfa Number of Sites	0.81

Since our district-level aggregates do not differentiate between refugees and host communities, we were unable to regress VASyR's refugee distribution against a refugee CDR population.

(b) Facebook

Figure 39 shows how Facebook Monthly Active Users (MAU) counts compare to the number of registered Syrian refugees. Beirut Governorate seems to be an outlier as it has a much larger number of users from the refugee community than the number of registered refugees; one possible reason could be the presence of non-Syrian Arabic language expats in Beirut.

Figure 39. Registered Syrian refugees distribution per district vs. FB estimates

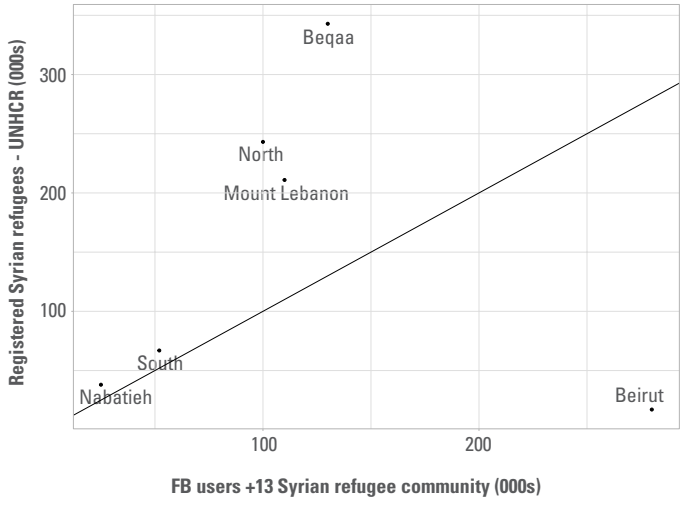


Table 7 displays the correlations between the estimated number of Facebook users from the refugee community and the number of registered Syrian refugees. When we exclude the suspected outlier Beirut, there is a strong positive correlation observed for the 5 governorates.

Table 7. Pearson correlation between FB population and VA SyR population distribution

Geographical disaggregation	Pearson correlation
All governorates	-0.093
Excluding Beirut	0.966

2. Economic activities

(a) CDRs

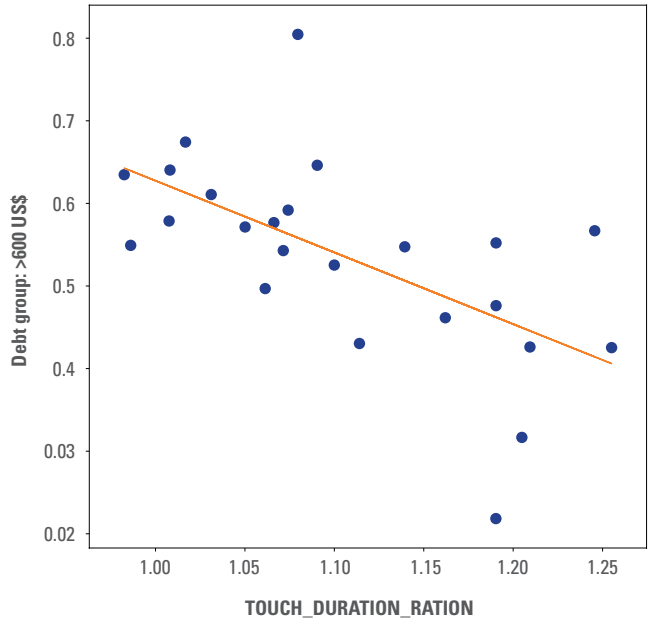
Table 8 shows correlations between the mean debt per household and per capita and the annual mobility indicator. We find a strong negative correlation for each of the variables and the mobility indicator of 72 per cent, which shows that the higher the mean debt per capita or mean debt per household in a district, the lower the mobility in that district, signifying that more people are leaving the district than moving into the district (negative mobility).

Table 8. Pearson correlation between CDR annual mobility indicator and VASyR economic indicators

Indicator	Pearson correlation
Mean Debt per Capita	-0.72
Mean Debt per Household	-0.72

Figure 40 shows the linear regression line of the CDR call duration ratio indicator used to predict the VASyR percentage of families with debt > $600.

Figure 40. Relationship between Syrian refugee household debt over $600 and call duration ration (out/in)



R^2 Score: 0.36

Mean Squared Error: 0.009

Table 9 shows the results in terms of R^2 score and mean squared error (MSE) of multiple regressions using the **K-Best CDR Features**, and their significance, to predict some of VASyR and CASLFS economic labels:

Table 9. CDR regression results for the economic target variables

Source	Target Variable	K-Best Features		R^2 score	MSE
		Feature	Relevant?		
CASLFS	Labour force participation rate (percentage)	**CDR** mobility	Yes	0.54	6
		CDR mobile data consumption download-to-total ratio	Yes		
		CDR ratio of number of outgoing to incoming calls	No		
VASyR	Percentage of families with debt > $600	**CDR** ratio of number of outgoing to incoming calls	Yes	0.35	0.009
		CDR ratio of outgoing to incoming call duration	Yes		

Table 10. Soundness test results

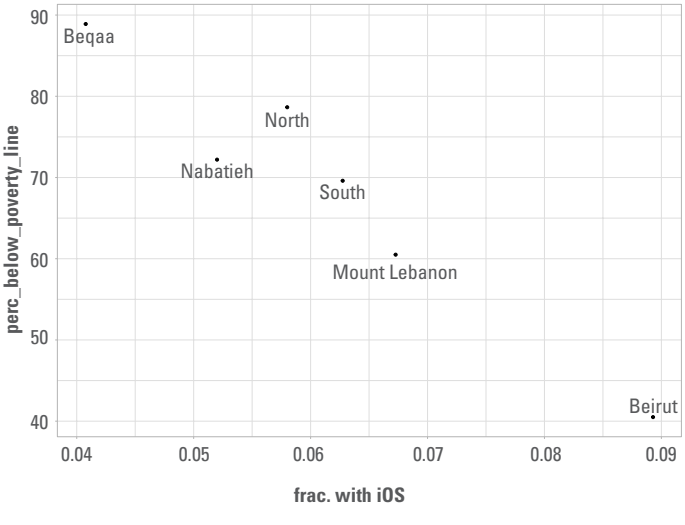
Target Value	Number of features	R^2 score using CDRs	R^2 using randomly generated data
CASLFS Population distribution	1 (linear regression)	0.9	0.01
VASyR perentage of families with debt > $600	1 (linear regression)	0.36	0.009

Soundness Test

We perform a soundness test to verify the effectiveness of the calculated CDR indicators and the models tested. Soundness tests analyse model uncertainty by comparing a baseline model to the model at hand. This is done by replacing the calculated indicators values with randomly generated but normally distributed and weakly correlated data, followed by repeating the same feed-forward, cross-validated feature selection and model fitting and testing. In this section, we compare the above results with the results of the soundness checks performed to calculate the same target values. To ensure the correctness of the soundness tests performed, we repeat these steps 1000 times and report the mean R^2 score for the 1000 runs.

Table 10 shows that despite the small sample size, the R^2 scores using the CDR indicators are much higher than those using random samples, with differences of up to 80 percentage points. This shows that the data and the indicators used in this work add valuable information and are likely not the result of spurious correlations.

Figure 41. Percentage of below poverty line vs. fraction using iOS



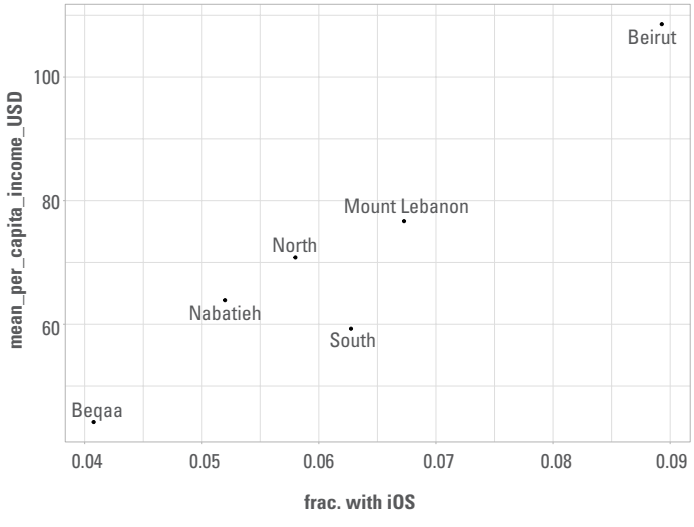
R^2 score: 0.89

Mean Absolute Error (MAE): 4.03

(b) Facebook

Figure 41 and 42 show the scatterplot of VASyR's percentage below poverty line vs. Facebook data on device/network type.

Figure 42. Mean per capita income vs. fraction using iOS



R^2 score: 0.8

Mean Absolute Error (MAE): 6.96

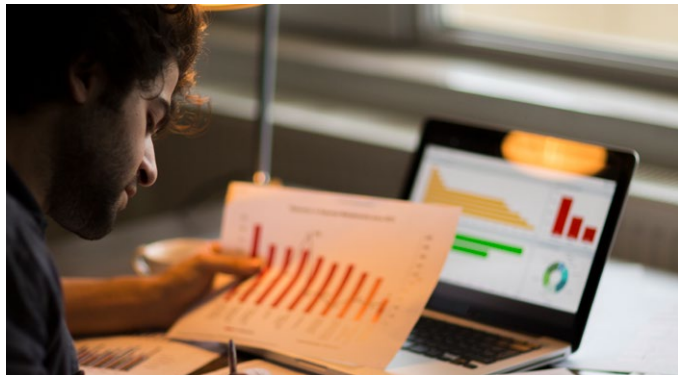


Table 11. FB regression results for the economic target variables

	Percentage of below poverty line	per capita income
Percentage of iOS device users	0.867	0.733
Percentage of Wifi users	0.6	0.067
Percentage of high-end phone (iPhone/Galaxy) users	0.6	0.2

Table 11 shows the Kendall correlation between the above Facebook indicators and VASyR labels, with a strong positive correlation between the percentage of iOS device users and each of the economic target variables.

3. Security

CDR

Table 12 shows the correlations between the population indicators based on CDRs, and tension with communities as a reason leading households to move out. We find a strong positive correlation between CDR indicators and VASyR security indicators, irrespective of the source of the CDR.

Table 12. Pearson correlation between CDR indicators and VASyR security indicators

	Pearson correlation
Alfa population indicator	0.7
Touch population indicator	0.65

Soundness tests analyse model uncertainty by comparing a baseline model to the model at hand. The data and the indicators used in this work add valuable information and are likely not the result of spurious correlations

Discussion

The proof of concept showed the potential of disparate data sources to predict the demographic and economic characteristics of communities despite the high levels of data aggregation.

In this pilot project, different non-traditional data sources were harnessed, particularly Mobile CDRs and Facebook Ad Platform, to serve as indicators relevant to 4 sustainable development goals (SDGs):

- Goal 1: End poverty in all its forms everywhere;
- Goal 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all;
- Goal 10: Reduce inequality within and among countries;
- Goal 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels.

Facebook Ad Platform's was able to estimate the percentage of Syrian refugees living in Lebanon below the poverty line and their mean capita per income. Mobile CDR indicators were able to predict multiple indicators relevant to the economic abilities and employment status of Syrian refugees and host communities, ranging from Syrian refugees' formal vs. informal work, percentage of Syrian refugees not participating in education or employment, host community labour force participation rate and more. The ability of these indicators to estimate traditionally collected poverty and employment indicators would likely provide policymakers with insights into the current status of their communities towards achieving these SDGs.

It is important to note that accessing data, particularly Mobile Call Detail Records, can be particularly challenging. It is a novel and innovative approach in Lebanon that carries with it several questions about privacy and data governance. For this project, the collaboration between UN ESCWA and the Central Administration of Statistics was critical for ensuring the accountability necessary for the Ministry of Communications to grant access to the information. Additionally, the number of geographic regions and level of aggregation in the data was limited to ensure user privacy. As mentioned throughout the document, there are other more ways to guarantee privacy while also providing more information. The expectation is that the present project will serve as a building block to consolidate the trust and necessary mechanisms to further explore the usefulness of this data source in analysing the living conditions of Syrian Refugees and Host Communities.

One of the difficulties that the UN and other agencies face in analysing the conditions of refugees is the associated cost and effort. Normally, data used to assess Syrian refugees' locations and needs are gathered through official government data and the UNHCR register of refugees through their data collection partners. Therefore, as stated in the Vulnerability Assessment of Syrian Refugees (VASyR) report, there is a consistent gap of Syrian refugees who have never reported to UNHCR.[61] As an example, according to the 2019 VASyR only 44 per cent of total families eligible for multipurpose cash assistance were provided with help.[62] In this sense, data gathered from non-traditional data sources such as those discussed in this project would be key to improving how

international and local agencies identify populations and their movements, so that they can provide services for Syrian refugees in Lebanon.

An understanding of population distributions is among the bases for planning the financial and human resources that need to be deployed geographically for crisis response. In order to analyse potential changes in refugee populations that could be detected through CDRs, the same months of CDR data were requested as in the VASyR data collection (April, May and June). Both the population indicator from CDRs and the population changes from VASyR revealed a decrease between 2017 and 2019, although the CDR indicator decreased at a faster rate. The advantage of using ground truth data for analyses is that weights can be assigned to call behaviour, improving the capacity of CDRs in predicting population changes, thereby adjusting for the different rates at which calling behaviour and population changes occur. For instance, in our univariate linear model, highly aggregated Call Detail Records with simple features (the number of calls) were shown to be good predictors of population proportions at the district level ($R^2 = 0.9$). The correlations between the number of calls made and the population were also high and positive, ranging between 0.88 and 0.99. When looking at the proxy for Syrian refugees, excluding Beirut, Facebook audience estimates correlated well with population proportions at the governorate level ($R^2 = 0.96$). While it is important to recognise the small number of data points available for

the estimation, these two data sources show potential for monitoring changes in the population distribution across the areas of study.

These findings are consistent with studies that, at a higher level of granularity, have shown the potential of these data sources for demographic analyses. It is expected that by increasing the data points available for the analysis, a better assessment of this potential can be made for Lebanon. This opportunity becomes even more relevant, since very soon the ground truth data for the Lebanese population (CASLFS 2019) can become less representative of the distribution of people around the country. Analysing CDRs could help monitoring changes in people's distribution, which can guide resource allocation. Even if this data source is not expected to provide an exact picture of people's distribution across space, population movements detected through it could indicate if an important movement in the population happened across the space which could lead to a reassessment of resource allocation.[63]

On the other hand, mobile phone usage indicators were good predictors of socioeconomic indicators of host communities and refugees, but to a lesser extent. When looking at economic indicators of Syrian refugees, the regression of family groups with debts greater than $600 with the number of calls and their duration has an R^2 of 0.35, meaning that the duration of calls explains 35 per cent of the variation in the number of debt groups. We expected this R^2 to be lower than that of other regressions, since the ground truth label referred to Syrian refugees only, while call durations represent both Syrian refugees and host communities. By contrast, regressing the labour participation from CASLFS on outgoing calls, mobility and data consumption returned a higher R^2 of 0.54, which is consistent with the units of analysis being more aligned between the ground truth label and CDR indicators. Similarly, Facebook audience estimates of the proxy for the refugee population had higher R^2 values. For instance, when measuring the percentage of people living under the poverty line, the fraction of individuals having an iOS device had an R^2 of 0.89, which is consistent with similar research by QCRI in other countries. Furthermore, the fraction of iOS device users explained 80 per cent of the variance in mean per capita income from VASyR at the governorate level.

One of the explanations for the lower explanatory power in the economic models is that having access to a phone introduces bias to the representativeness of call behaviour for vulnerable populations, meaning that the most vulnerable are underrepresented. However, despite the lower predictive power than in population predictions, the R^2 scores were still relatively high. This makes sense, since the lower number of calls, call duration and internet use at the district level can be indicative of a greater presence of vulnerable populations. In other words, people residing in less well-off areas tend to have less access to mobile phones, which can then translate into lower numbers in the indicators analysed.[64] Furthermore, even among those who have access to phones, a lower capacity to recharge the bundles can be indicative of lower economic resources.

Among the key advantages of these innovative data sources in this context is their capacity to be updated at a higher frequency than traditional data sources. This can be helpful in the context of crises, where decisions often need to be made with little to no recent information on the populations concerned under pressing time constraints. To this end, key events in the areas of study can be analysed, to understand the extent to which these data sources provide insights on sudden changes in population distributions through mobility. In the context of this study, our focus on April, May and June meant that events taking place outside of these boundaries could not be directly analysed. Nonetheless, by comparing the three-month average from one year to another, we found a 5 per cent decrease in Bcharre's population indicator, which happened to coincide with the civil conflict that led to Syrian refugee evictions in the area. While this analysis is not meant to provide causal evidence, it is an area worth exploring in more detail with additional data, particularly because CDRs have been employed in other contexts to detect population movements after natural disasters, like the earthquake in Nepal.[65]

Overall, this study shows the potential of disparate data sources to predict the demographic and economic characteristics of Syrian refugees and Host Communities in Lebanon. Despite the high levels of data aggregation and focus on only two governorates, the models were able to explain between 50 and 90 percent of the variation in ground truth labels for Host Communities and Syrian Refugees. The main implication of this proof of concept is that further analysis, through greater granularity in the units of analysis, as well as both longer and more granular periods of time, can increase the capacity of the models to inform on sociodemographic changes of both populations. To this extent, granting access to this type of information would likely contribute quantitative evidence and insights into the variation of indicators tracked by official statistics in the medium term and provide quick insights for crisis response in the short term.

Lessons Learned and Recommendations

A. Enabling conditions crucial to project success

The completion of this project would not have been possible without the process of creating, alongside UN-ESCWA and CAS, the conditions, systems, governance standards, incentives and capacities for the systematic and safe access and analysis of non-traditional data sources. The stewardship of local and multilateral actors made possible the iterations of the research process and questions, data access and analysis. The significant political will, and strong partnerships developed in the framework of this project, made possible its completion. Furthermore, the CODE (Council for the Orientation for Development and Ethics) was a crucial body to ensure the appropriateness and soundness of the research, as well as its ethical standards.

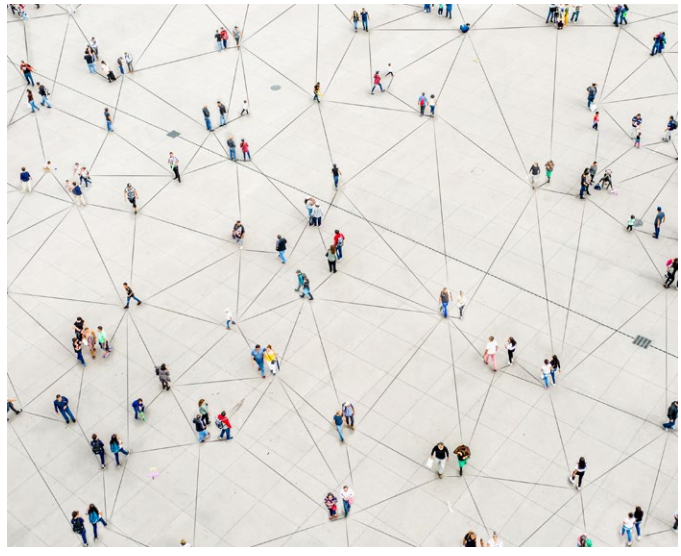
B. Privacy-conscientious design

Through the research design, particularly that of the Call Detail Records, the project was able to conduct thorough and safe research with strong privacy considerations, including on-site data processing, geographic and temporal data aggregation. Not only does this confirm that a privacy conscientious research design is possible – even with its trade-offs – but it also elicits opportunities to do further research, while keeping privacy as one of the key research focuses.

C. Accessing call detail records

CDRs can easily amount to terabytes of data on social behaviour, which has been shown to correlate with socioeconomic and demographic indicators. However, the scale of information that can be derived from them must be balanced with the protection of privacy. Among the recommended methods, data aggregation is one of the simplest and most useful. However, when aggregating data, it is fundamental to do it on the right indicators, so as to maintain group information while minimizing information loss. Since the "right" indicators can vary between research topics, two approaches are advised.

1. Maintain close communication between researchers and Telecom Operators. This implies that data requests should be specified from the beginning, but they can be verified throughout the process. For these purposes, Telecom Operators should be in the capacity of providing examples of their data structure, so that researchers can provide examples of code to create the indicators and aggregate them. This is particularly useful for contexts where Telcos are not used to collaborate on social data science research.



2. Utilise readily made tools to create hundreds of different indicators that are often used in CDR research. For instance, Bandicoot[66] is an open source library that allows the creation of commonly used indicators in a relatively easy way.

There is great potential to be reaped from integrating census or survey data to individual CDRs, as this opens the possibility of using the highest level of granularity in both sources and algorithms can learn to combine information from both sources. As explained in the CDR section, privacy can be preserved through cryptographic advancements such as private set intersections.

The significant political will, and strong partnerships made the project completion possible.

The "Council for the Orientation for Development and Ethics" was a crucial body to ensure the appropriateness and soundness of the research, as well as its ethical standards.

Annex

1. Combining different data sources

Combining indicators gathered from the different data sources can further enhance the model's ability to predict target values. However, a larger sample is required to make use of more indicators simultaneously. In the case of this project, due to the aggregated nature of the data at hand, using a complex model combining numerous features increased the risk of overfitting.

In this annex, we report the results of performing multiple linear regression while combining all our sources at the district level: CDRs, Facebook, and GDELT. In order to predict target economic values, we combined the CDR economic indicators along with the Facebook economic indicators discussed in the results section. In this combination, for CASLFS economic target values, we use Facebook's non-expat (host communities) indicators and drop the refugee indicators (and vice versa for VASyR economic target values). In order to predict the security target values, we combine the GDELT indicators along with the CDR mobility and population indicators. We also performed the same random sample-based robustness checks to test whether the model is overfitting. Table 13 shows the R^2 score of fitting a regression model of 3 or more features on our 12-row dataset, in addition to the R^2 score of the normalised random data for the robustness check. As the table shows, the difference between the two R^2 scores is not significant for most of the rows. This is due to the model overfitting which is caused by the small sample size.

Table 13. FB regression results for the economic target variables

Label	No. of features	R^2 of Combined Data	R^2 of Normalized random data (0,1)
Percentage of families < SMEB ($87)	3	0.51	0.62
Percentage of families' debt ≤ $200	3	0.6	0.62
Percentage of Syrian youth (15-24) who are not in education, not employed and not attending any training (NEET)	3	0.7	0.62
Percentage of families self-described as rich	3	0.63	0.61
Labour force participation rate	3	0.62	0.61
Of families that moved in the past 6 months, (per cent) reason: Tension with community or restrictive measures	3	0.71	0.7
Percentage of HH who had an incident with their current landlord in the past 6 months	3	0.74	0.7

Endnotes

1. For example, only the Ministry of Telecommunications and the Telecom operators are privy to the calls data in this study.
2. We were able to obtain indicators on the economic and social condition of refugees and host communities from this data, which correlate with official sources of statistics.
3. This can be evidenced by the pay gap between host and refugee communities despite the small difference in education level, or the informal work by refugees which was picked up by the phone calls data.
4. The mass eviction from the whole region in the North, which coincided with one incident in Zgharta, is a clear indicator of the need for better synchronization between leaders of refugee communities and host communities.
5. UNHCR, Lebanon Fact Sheet, 2020. Available at https://reliefweb.int/report/lebanon/unhcr-lebanon-factsheet-january-2020.
6. More information on D4R | Data Pop-Alliance.
7. https://gs.statcounter.com/social-media-stats/all/lebanon.
8. E. Zagheni, I. Weber, K. Gummadi, "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants" Population and Development Review, The Population Council, Inc., vol. 43, No. 4, pp. 721-734, Dec. 2017, doi:10.1111/padr.12102.
9. https://www.facebook.com/business/news/expat-targeting-uk.
10. https://developers.facebook.com/docs/marketing-api/audiences/.
11. J. Palotti and others, "Monitoring of the Venezuelan exodus through Facebook's advertising platform". PloS one, vol. 15,2 e0229175, Feb. 2020, doi:10.1371/journal.pone.0229175.
12. M. Fatehkia and others, "Mapping socioeconomic indicators using social media advertising data". EPJ Data Science, vol. 9, No. 22, July 2020, doi:10.1140/epjds/s13688-020-00235-w.
13. M. Fatehkia, R. Kashyap, I. Weber, "Using Facebook ad data to track the global digital gender gap, World Development", vol. 107, pp. 189-209, 2018, doi:10.1016/j.worlddev.2018.03.007.
14. https://cloud.google.com/bigquery.
15. https://developer.twitter.com/en/docs/ads/campaign-management/api-reference/features.
16. Examples include but not limited to: python packages https://pypi.org/project/GetOldTweets3/ and https://pypi.org/project/twitter-scraper/.
17. UNHCR, Vulnerability Assessment of Syrian Refugees in Lebanon, 2019.
18. CAS, Labor Force and Household Living Conditions Survey, 2019.
19. J. Blumenstock, G. Cadamuro, R. On, "Predicting poverty and wealth from mobile phone metadata", Science, vol. 350, No. 6264, pp. 1073-1076, Oct. 2015, doi:10.1126/science.aac4420.
20. V. Frias-Martinez, J. Virseda, "On the relationship between socio-economic factors and cell phone usage", ACM ICTD '12, pp. 76-84, 2012, doi:10.1145/2160673.2160684.
21. UN Global Working Group on Big Data for Official Statistics, The Handbook on the use of Mobile Phone data for official statistics, 2017. Available at https://unstats.un.org/bigdata/taskteams/mobilephone/MPD%20Handbook%2020191004.pdf.
22. UN Trade Statistics, Overview of the sources and challenges of mobile positioning data for statistics, 2014. Available at https://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf.
23. P. Deville and others, "Dynamic population mapping using mobile phone data", PNAS Early Edition, 2014, doi:10.1073/pnas.1408439111.
24. B. Furletti and others, "Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach", 47th SIS Scientific Meeting of the Italian Statistica Society, June 2014.
25. J. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda", Information Technology for Development, vol. 18, Feb. 2012, doi: 10.1080/02681102.2011.643209.
26. M.Bakker and others, "Measuring Fine-Grained Multidimensional Integration Using Mobile Phone Metadata: The Case of Syrian Refugees in Turkey", Guide to Mobile Data Analytics in Refugee Scenarios, pp. 123-140, Sep. 2019, doi:10.1007/978-3-030-12554-7_7.
27. More info https://www.touch.com.lb/autoforms/portal/touch/personal/prepaid/altawasol/overview.
28. V. Frias-Martinez, J. Virseda, "On the relationship between socio-economic factors and cell phone usage", ACM ICTD '12, pp. 76-84, 2012, doi:10.1145/2160673.2160684.
29. https://civilsociety-centre.org/sir/bsharri-residents-protested-evict-syrian-refugees.
30. https://www.washingtonpost.com/world/middle_east/all-lebanon-is-against-them-a-rape-murder-sours-a-country-on-its-syrian-refugees/2017/10/10/afa13010-a792-11e7-9a98-07140d2eed02_story.html.
31. https://civilsociety-centre.org/sir/bcharre-municipality-asks-syrian-refugees-leave.
32. https://data.worldbank.org/indicator/IT.CEL.SETS?locations=LB.
33. https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=LB.

34. https://reporting.unhcr.org/node/2520?y=2017#year.

35. https://reporting.unhcr.org/node/2520?y=2018#year.

36. https://reporting.unhcr.org/node/2520?y=2019#year.

37. https://www.worldometers.info/world-population/lebanon-population/.

38. UN Global Working Group on Big Data for Official Statistics, The Handbook on the use of Mobile Phone data for official statistics, 2017. Available at https://unstats.un.org/bigdata/taskteams/mobilephone/MPD%20Handbook%2020191004.pdf.

39. B. Sakarovitch, M. De Bellefon, P. Givord, M. Vanhoof, "Estimating the Residential Population from Mobile Phone Data, an Initial Exploration", Economie et Statistique/ Economics and Statistics, pp. 109-132, 2019, doi:10.24187/ecostat.2018.505d.1968.

40. F. Clarke, C. Chien, "Visualising big data for official statistics: The abs experience", Data Visualization and Statistical Literacy for Open and Big Data, pp. 224-252, 2017, doi:10.4018/978-1-5225-2512-7.ch009.

41. P. Daas, M. Puts, B. Buelens, P. Hurk, "Big Data as a Source for Official Statistics", Journal of Official Statistics, vol. 31, No. 2, pp. 249-262, 2015, doi:10.1515/JOS-2015-0016.

42. X. Lu and others, "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh", Global Environmental Change, vol. 38, ISSN 0959-3780, pp. 1-7, 2016, doi:10.1016/j.gloenvcha.2016.02.002.

43. J. Novak, R. Ahas, A. Aasa, & S. Silm, "Application of mobile phone location data in mapping of commuting patterns and functional regionalization: A pilot study of Estonia", Journal of Maps, vol. 9, pp. 10-15, 2013, doi:10.1080/17445647.2012.762331.

44. V. Frías-Martínez, C. Soguero, E. Frias-Martinez, "Estimation of urban commuting patterns using cellphone network data", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 9-16, 2012, doi:10.1145/2346496.2346499.

45. G. Pestre, E. Letouzé, E. Zagheni, "The ABCDE of Big Data: Assessing Biases in Call-Detail Records for Development Estimates", The World Bank Economic Review, vol. 34, No. 1, pp. S89-S97, Feb.2020, doi:10.1093/wber/lhz039.

46. B. Pinkas, T. Schneider, O. Tkachenko, A. Yanai, V. Rijmen, "Efficient Circuit-Based PSI with Linear Communication". Advances in Cryptology – EUROCRYPT 2019, vol. 11478, pp. 122-153, 2012, doi:10.1007/978-3-030-17659-4_5.

47. E. Zagheni, I. Weber, K. Gummadi, "Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants" Population and Development Review, The Population Council, Inc., vol. 43, No. 4, pp. 721-734, Dec. 2017, doi:10.1111/padr.12102.

48. S. Spyratos, M. Vespe, F. Natale, I. Weber, E. Zagheni, M. Rango, "Quantifying international human mobility patterns using Facebook Network data", PLoS One 2019, vol. 14, No. 10, Oct. 2019, doi:10.1371/journal.pone.0224134.

49. A. Dubois, E. Zagheni, K. Garimella, I. Weber, "Studying Migrant Assimilation Through Facebook Interests", International Conference on Social Informatics, SocInfo 2018, pp. 51-60, Sep. 2018, doi:10.1007/978-3-030-01159-8_5.

50. D. Galla and J. Burke, "Predicting Social Unrest Using GDELT", Machine Learning and Data Mining in Pattern Recognition MLDM 2018, vol. 10935, July 2018, doi:10.1007/978-3-319-96133-0_8.

51. F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, H. Wang, "Predicting Social Unrest Events with Hidden Markov Models Using GDELT", Journal of Discrete Dynamics in Nature and Society, vol. 2017, pp. 1-13, 2017, doi:10.1155/2017/8180272.

52. S. Bertoli and others, "Integration of Syrian Refugees: Insights from D4R, Media Events and Housing Market Data", Sep. 2019.

53. I. Abu Farha, W. Magdy, "Mazajak: An Online Arabic Sentiment Analyser", Proceedings of the Fourth Arabic Natural Language Processing Workshop, ACM, pp. 192-198, doi:10.18653/v1/W19-4621.

54. Amazon Mechanical Turk, https://www.mturk.com/.

55. M. Imran, Muhammad, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Patrick, "Extracting Information Nuggets from Disaster- Related Messages in Social Media", The 10th International Conference on Information Systems for Crisis Response and Management ISCRAM, May. 2013.

56. UNDP, Analyzing Refugee-Host Community Narratives on Social Media. Available at https://data2.unhcr.org/en/documents/details/69992.

57. N. Öztürk, S. Ayvaz, "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis", Telematics and Informatics, vol. 35, No. 1, pp. 136-147, Apr 2018, doi:10.1016/j.tele.2017.10.006.

58. UNHCR, Social Media and Forced Displacement: Big Data Analytics & Machine-Learning, 2017. Available at https://www.unhcr.org/innovation/wp-content/uploads/2017/09/FINAL-White-Paper.pdf.

59. M. Imran, Muhammad, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Patrick, "Extracting Information Nuggets from Disaster- Related Messages in Social Media", The 10th International Conference on Information Systems for Crisis Response and Management ISCRAM, May 2013.

60. Since our variables are both proportions, we acknowledge that a linear regression is not suited to model the data, and that instead logistic regression should be deployed for instance. However, given the small sample size, the estimates for the logistic regression will be unreliable as they will be biased (see for example Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). "Bias in odds ratios by logistic regression modelling and sample size". BMC medical research methodology, vol. 9, pp. 56-60). Some researchers argue that the linear regression could serve as a rough approximation as it is easier to interpret than the logistic regression and produces acceptable results (see for example Ottar Hellevik (2009) "Linear versus logistic regression when the dependent

variable is a dichotomy", Quality & Quantity, vol. 43, pp. 59-74). Taking all of these points into consideration, we proceed with the linear regression for simplicity and illustrative purposes, as our overarching aim is to only determine whether a relationship exists between the variables.

61. UNHCR, Vulnerability Assessment of Syrian Refugees in Lebanon, 2019.

62. UNHCR, Vulnerability Assessment of Syrian Refugees in Lebanon, 2019.

63. https://web.flowminder.org/publications/rapid-and-near-real-time-assessments-of-population-displacement-using-mobile-phone-data-following-disasters-the-2015-nepal-earthquake.

64. Contextually, people tend to opt for internet use through apps like WhatsApp instead of regular calls. The inclusion of mobile internet usage in this project was used to capture these changes and improve predictions of socioeconomic indicators.

65. https://web.flowminder.org/publications/rapid-and-near-real-time-assessments-of-population-displacement-using-mobile-phone-data-following-disasters-the-2015-nepal-earthquake.

66. Bandicoot, "an open source Python toolbox to analyze mobile phone metadata", found at https://cpg.doc.ic.ac.uk/bandicoot/.

20-00449