Mobile data challenges for human mobility analysis and humanitarian response[1]

Albert Ali Salah
Utrecht University
Social and Affective Computing
Dept. of Information and Computing Sciences
a.a.salah@uu.nl
http://orcid.org/0000-0001-6342-428X

Telecommunication operators have a unique perspective on human mobility; they know the locations of their customers, for most of the time. In recent years, a number of initiatives were organized by telecommunication operators in which mobile call data records (CDR) were carefully anonymised, aggregated, and opened to researchers in form of a challenge for providing new insights into people's movements and displacement patterns. An example is the Data for Refugees Challenge, organized with the expressed aim "to improve the living conditions of over 3.5 million Syrian refugees in Turkey". Typically, the mobile phone datasets collected by a telecommunication operator only include movement patterns observed within a single country, but with additional assumptions, it is also possible to gain insights into movements across borders. This chapter provides an historical overview of research on mobility analysis through mobile CDR, highlights practical issues such as data gaps and biases, discusses ethics and privacy principles that must be taken into consideration when working with such sensitive data, and argues that migration studies and humanitarian response projects may benefit greatly from the use of real-time or historical mobile CDR data.

Keywords: Mobile CDR, migration studies, refugees, data science, computational social science, AI for social good

<1> Introduction

The capabilities of computer systems expand every day and new approaches are being developed for computer analysis of human behaviour (Salah and Gevers, 2011). At the largest scale of modeling, data collected over long periods of time from millions of people living in a region can be visualized, analyzed, and interpreted with the help of computer systems. Especially during the COVID-19 pandemic, the value of large scale behaviour modeling and analysis became apparent, both for prediction of the spread of the disease, as well as in assessing measures in its control (Oliver et al., 2020).

A common processing pipeline starts with 'sensing' the behaviours in question, using physical sensors like cameras, or virtual sensors like user traces left on the Internet and on phone-based applications. The analysis of spatio-temporal data generated by these sensors can use models developed by social scientists, but also be accomplished in a data-driven way, with the help of statistics, pattern recognition, and machine learning. Finally, the results of the

[1] This is the uncorrected author proof. Please cite as: Salah, A.A., "Mobile data challenges for human mobility analysis and humanitarian response," in M. McAuliffe (ed.) Handbook of Migration and Technology, Edward Elgar, forthcoming.

analysis, which can include prediction of behaviours, or explanation of their origin or dynamics, are put to some use. They can feed applications that provide better services, or help them accomplish their tasks more effectively.

Computational social science is a new field that "leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviours" (Lazer et al., 2009 p. 722). Using social media or mobile phones, it is possible to get data from millions of people in a very short amount of time. It is also possible to observe longitudinal changes as reflected on such data sources, because the measurements can be repeated regularly. For large-scale human behaviour analysis, traditional methods of data collection, such as surveys, interviews, and focus groups have limitations, including the high cost of data collection, response biases, and the difficulty of ensuring the same conditions over longitudinal designs (Margolin et al. 2013). The difficulties are exacerbated in the case of studying migration and human mobility, which can greatly benefit from the analysis of new data sources (Zagheni and Weber, 2015).

In this chapter, we discuss a specific sensing modality for large-scale human mobility analysis, namely, mobile phone data, that may overcome some of these limitations. In particular, mobile phones are ubiquitous in many countries (and increasingly used, thus implying less bias by wealth)[i], and offer very high temporal resolution in data acquisition. The potential of mobile phone data for the analysis of human mobility is perhaps not very surprising, as the location of a mobile phone can be tracked with great accuracy. Additional information from a smart phone and its applications can greatly increase this potential.

This power comes with a range of ethical and privacy issues that need to be considered when dealing with large scale, sensitive data analysis. In particular, proprietary data sources, such as call detail records of mobile telecommunications companies, are extremely valuable, but also very sensitive data collections. In this chapter we discuss some mechanisms to share and analyze such collections responsibly.

This chapter starts with describing how mobile data are used for behaviour analysis in general, and mobile call data records in particular. We then explain the concept of data collaboratives, and summarize efforts in several large-scale mobile data challenges, including the recent Data for Refugees Challenge (Salah et al., 2019). We discuss mobile data in the context of mobility, migration and displacement, and for humanitarian response. We deal with the ethical and privacy issues at some length and we conclude the chapter by discussing key implications for research and governance.

<2> Mobile Phone Data for Behavior Analysis

The potential of using mobile phone data in gaining in-depth insight into movements of people was revealed forcefully during the COVID-19 pandemic (Oliver et al. 2020), but investigations of the potentials and limits started much earlier. In a landmark study, Eagle and Pentland (2006) provided 100 smart phones to students at MIT, and just by tracking simple indicators, such as call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle), were able to obtain interesting insights into the social behaviours of the students[ii]. Their goal was to explore the capabilities of smart phones for enabling social scientists to investigate human interactions beyond the traditional survey or simulation based methodologies. Indeed, applying well-known pattern recognition techniques on this type of data revealed most frequented locations of individuals, their social contacts (by using regular proximity with other individuals), networks and groups, and even organizational dynamics. Given data from a new student, they were able to classify whether the student was a first year student or a graduate student (with over 95% accuracy), and given half a day of observing a student's behaviour, they could predict the rest of the day's behaviour with almost 80% accuracy (Eagle and Pentland, 2009). (Bayir et al. 2009) found that users in this dataset spent 85% of their time in three to five favourite locations. One can argue that a small number of MIT students may not have a lot of variability in their daily routines, yet daily routines and social activities are important to many people, and these studies predicted that such methods will play an increasingly important role in behaviour research.

What was also initially surprising was how much information could be extrapolated by combining such mobile phone data with other data sources. In 2005, Ahas and Mark combined mobile phone data with demographics and survey data in what they termed the 'Social Positioning Method,' for the analysis of social spaces in Estonia. They remarked that one can study social flows in space and time through such an approach, but questions of privacy and freedom of the individual should be asked.

Mobile applications can collect a lot of information from their users, but they are restricted to the people who use the application, whereas data available to telecommunication operators encompasses all cellphone users. In the remainder of this section, we focus on such data.

<2.1> Mobile Call Data Records

A 'call data record' (CDR) is a data structure that contains the exact time, duration, and location (antenna) of a call, as well as the calling number and callee number. If within the same telecommunications operator, antenna IDs of both sides of the call can be known. A similar record will be created for the SMS messages. Naturally, the actual content of the messages or calls are not stored; this is mostly a record that the telecommunication company (or, equivalently, the network operator) creates for billing and accounting purposes. They may also be internally used to optimize resources and infrastructure.

What would count as a personal information in a CDR database is the phone number itself, and removing this number, or replacing it with a random indicator could be a way of anonymizing such data. However, this by itself is not nearly enough, and there are additional considerations to be taken into account. If there are too few people around an antenna at a given time, the CDR can be associated to a known person by a process of elimination. Subsequently, spatial and temporal aggregation is used in addition to anonymization (see Figure 1). Furthermore, external data sources can help de-anonymizing such a dataset. For example, if precise information for a number of travels of a person is known, that person can be identified in such a database. One solution is to track individuals for very short periods, such as for one or two weeks at a time. These records are called *tracklets,* and each time a person is sampled from the pool of users, they are given a new random identifier. When we just look at tracklets from thousands of people, we can see the dynamics of mobility between connected locations without jeopardizing the anonymity of individuals from which the tracklets were obtained.

[Insert Figure 1 about here]

Regularities in people's living allow getting semantically rich information out of CDR records. By determining the most frequently used locations for phone calls during work hours, it is possible to estimate the location(s) where a person works. The same can be done for a home location by looking at early morning or late night calls. Once these landmarks are known, they can serve as anchors for classifying behaviour.

<2.2> Sharing mobile CDR

Detailed mobile CDR is a rich data source for looking at mobility, but it is generated and used internally by telecommunications operators. Since it is highly sensitive, it is not released to the public. These datasets are analyzed from time to time by a small group of researchers who are granted access through project collaborations (Deville et al., 2014; Lu et al., 2016) or by governments, under special regulations. According to Maxmen (2019), more than 20 mobile phone companies participated in such efforts. This includes operators in 100 countries that backed the Big Data for Social Good initiative, sponsored by the Global System for Mobile Communications Association (GSMA), which also developed an online resource to facilitate further use cases[iii].

There are in fact several data sharing models for CDR, each with different strengths and weaknesses. Letouzé and Oliver (2019) list five models, based on an analysis of scale and protection level (de Montjoye et al., 2018):

1. Limited data release: Private companies package the data and release it to a small number of groups. Scale and protection are both low.

2. Remote access: Company grants access for limited data to a few authorized groups. The scale is similarly low, but protection level is higher.

3. Access through application programming interfaces (API)s: This is a question and answer model, where data users send a query to an API, and data are aggregated within the private company, never leaving the premises. Only the 'answer' is sent out. It is both scalable and secure, but difficult to set up and maintain. A good example is the Open Algorithms (OPAL) framework, which is currently being tested in several countries (Letouzé, 2019).

4. Pre-computed indicators: Data are not shared; only indicators are computed and shared. This limits the analysis and precludes innovative approaches for using the data. Similarly, deriving statistics from the data and producing a synthetic dataset for sharing will miss aspects not included in the statistical modeling already.

5. Data collaboratives or cooperatives: Data cooperatives are institutional agreements for incentivizing data sharing, where individuals will pool their data under one umbrella, and have control (and sometimes revenue) when sharing data.

Verhulst and Young (2019, p.465) remark that the task of understanding the causes and consequences of population movements can partly be addressed "through the targeted analysis of datasets dispersed across stakeholders in governments, the private sector, and civil society." For this purpose, they define the concept of a *data collaborative*, which is a collaboration between actors from different sectors, both public and private, enabling improved data collection, sharing and usage. This is a data-centric approach to mobility and migration, and assumes that information-related problems are of utmost importance for these fields. Through data collaboratives, it becomes possible for the stakeholders to access more relevant, real-time and detailed data, which in turn guides decision and policy making. For migration and mobility specifically, there are currently several projects underway aiming to use new data sources via such collaboratives. For example, the SoBigData and HummingBird projects funded by European Union's Horizon2020 program are investigating data collaborations (including mobile CDR) related to migration policy and migration indicators[iv].

<2.3> Mobile Data Challenges

In the rest of this section, we describe several initiatives to open CDR data to the scrutiny of a large number of researchers as data collaboratives with telecommunication operators. Each of these initiatives required significant effort by a large number of people addressing technical, organizational, and ethical aspects of such a data collaborative, as well as the cooperation and commitment of the telecommunication operator in question. There is also a strong, interdisciplinary community building effort around these topics. The Analysis of Mobile Phone Networks Conference (NETMOB), initiated in 2010, continues to be a gathering place for scholars who explore the potential of mobile phone data in various settings[v].

The first large initiative involving a data collaborative with mobile data was the Data for Development (D4D) Challenge, which was initiated by Orange Cote D'Ivoire in 2012 (Blondel et al., 2012). This initiative opened a big dataset of mobile CDR for the first time to a large group of scientists, in the form of a data challenge. The data consisted of five months of anonymized CDR records and antenna positions, and the aim of the challenge was "to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Ivory Coast population" (Blondel et al., 2012, p.1).

The CDR data for D4D contained:

- antenna-to-antenna traffic on an hourly basis,
- individual trajectories for 50,000 randomly sampled users for two week time windows with antenna location information,
- individual trajectories for 500,000 randomly sampled users over the entire observation with sub-prefecture location information,
- a sample of communication graphs for 5,000 customers.

The scientific challenge, which ran for eight months, required institutions interested in the data to sign a legal agreement with binding terms and conditions, and the proposals were screened by an evaluation committee. 263 project proposals were submitted to the challenge, and over 80 reports were received.[vi]

Some of the outcomes of the projects were potentially useful in improving the infrastructure in Ivory Coast. For example, an analysis of the transit network jointly with CDR revealed the differences of Origin-Destination flows indicated by the CDR with the transit network, and four new routes were suggested for addition to the city transit network, decreasing the citywide travel time by 10% (Di Lorenzo et al., 2015). Other project outcomes suggested improvements for the healthcare system through modeling of disease dynamics, provided better population statistics and economic indicators, and provided insights into the social dynamics of Ivory Coast.

In 2014, Telecom Italia opened a dataset of its own CDR records from two cities as a challenge[vii]. In the same year, Sonatel Senegal and Orange Labs embarked on a second edition of D4D, by providing CDR collected over an entire year from over nine million unique mobile phones (de Montjoye et al., 2014). By that time, it was already known that four spatio-temporal points are sufficient to find a person in a mobility database of 1.5 million users with 95% accuracy (de Montjoye et al., 2013), so the data were aggregated and anonymized to prevent such re-identification. The same data aggregation scheme was used in 2018, during the Data for Refugees (D4R) Challenge, which we describe in the Section 3.3 in more detail.

The ethical safeguards and privacy measures used in all these challenges focused on the principle of "do no harm", where responsible data practices are promoted, a broad set of stakeholders involved in decision process (for example in the Data for Refugees Challenge, this included refugees and institutions protecting the rights of refugees), and measures are taken to minimize the risk to individuals (Salah et al., 2018, Vinck et al., 2019). We discuss these issues in Section 4 on ethics and privacy, and broadly comment on existing frameworks and concerns.

<3> Mobility, migration, displacement analysis and humanitarian response with CDR

In this section, we provide examples on how mobile CDR can be used for the analysis of mobility, migration and displacement, discussing specific applications, including analysis of refugee movements and humanitarian response.

<3.1> Movement patterns

Movement patterns of the masses show great regularity, and inspecting regularities and breaks from regularity is a great way to start analysis of mobile CDR. Dobra et al. (2015) processed mobile phone records from Rwanda with an unsupervised learning approach to detect deviations from expected behaviour, and connected the identified anomalous days and locations with records of violent and political events such as protests, and natural disasters like earthquakes. In a similar approach, Gundogdu et al. (2016) analyzed CDR data from Ivory Coast and showed that movements of different ethnic groups on religious holidays were different. It was possible to use regular visits to a basilica during Christian holidays as a proxy indicator for religion, allowing the profiling of users in terms of this variable[viii]. Even though such a classification will be noisy, when hundreds of thousands people are used in the analysis, it can produce meaningful patterns. In a sense, computational social science replaces the carefully controlled experimental settings of traditional social science with a much less controlled setting, but relies on orders of magnitude larger sample sizes to reduce the effect of noise and the uncontrolled variables (Giannotti et al., 2016).

The CDR collected from an entire country gives a detailed picture about internal displacement and mobility, connecting not only cities to cities, but potentially giving prefecture-level mobility patterns. But how can we deduce cross-border movements? One obvious solution is to work with multiple telecom operators at the same time, and link roaming and cell phone usage across operators. This, however, may create serious privacy concerns. In some cases, the context of mobility may provide cues. In the Ivory Coast study mentioned above, the data were collected during a severe civil conflict, and many people escaped the country over borders. When CDR is tracked longitudinally, it is possible to see the movement tracks leading to a border and the cessation of activity for the remainder of the data collection period, for a number of users. These tracks, with a high probability, illustrate displacement of people across borders.

Gonzáles et al. (2008) tracked trajectories of 100.000 anonymized mobile phone users, and observed a high degree of temporal and spatial regularity. Mathematical models were used to characterize the distribution of movements of individuals. Most individuals moved within a small radius of gyration, which denotes the distance travelled by a person when observed for a specific time. It is clear that internal (or external) migration and displacement will create quite different trajectories, which can be modeled similarly.

Ahas et al. (2018) classified people who were traveling abroad from Estonia into groups of tourists, commuters, transnationals, and foreign workers, using roaming data from mobile operators, depending on the number of visits and the number of days spent in destination. Since register-based data is not very suitable to describe the transnational community, mobile data provides a good alternative.

Generalized radiation models are the state of the art in modeling diverse mobility scenarios (Kang et al., 2015). Isaacman et al. (2018) incorporated weather conditions into an extended radiation model to predict mobility in La Guajira, Colombia during a drought, and used mobile CDR as a ground truth to contrast different models. They conclude that mobile CDR is powerful for modeling climate change related migration. Both gravity and radiation models were developed with geospatial population information as their input sources. If one has access to mobile CDR, models can use additional information. For example, Palchykov et al. (2014) developed a simple communication model based on the frequency of mobile phone calls between two locations and their geographical distance.

<3.2> Indicators

The use of mobile CDR is not only a good source for predicting movements, but also for linking mobility behaviours to other indicators, which are potentially even more powerful. It is possible to derive proxies from mobile CDR for many indicators, including population density (Deville et al., 2014), urban activity (Reades et al., 2007), travel behaviour (Wang et al., 2019), civic engagement and political participation (Campbell and Kwak, 2010), economic status and poverty (Šćepanović et al., 2015), social integration (Bakker et al., 2019) and undeclared employment (Bruckschen et al., 2019). The raw data is first aggregated and converted into more informative representations, such as origin-destination matrices, total travelled distance for individuals or geographical areas, as well as spatial and temporal properties of routine behaviours (Hughes et al., 2016). Regularities of movement lead to tagging locations of 'home' and 'work' for each individual (Isaacman et al., 2011). Further analysis on social network, travel behaviour and frequently visited places, makes behavioural profiling very powerful.

An example is predicting poverty and wealth from mobile data, for example by measuring commuting patterns (Šćepanović et al., 2015). Some qualitative information about the socio-economic status of different regions needs to be combined with mobile CDR in this case. Blumenstock et al. (2015) used data from Rwanda’s largest mobile phone network for wealth analysis, and verified their findings with follow-up phone surveys of a geographically stratified random sample of 856 individual subscribers. When they aggregated the results, the predicted composite wealth index, computed from 2009 call data and aggregated by administrative districts showed remarkable similarity with the actual composite wealth index as computed from a 2010 government Demographic and Health Survey (DHS). The latter was collected from about 12.800 households, whereas the call data was obtained from 15 million individuals over one year.

The most alluring benefit of using mobile CDR is arguably the cost of obtaining information. The cost of a national household survey is estimated at over $1 million, taking 12 to 18 months to complete (Jerven, 2014), but the CDR can be accumulated and analyzed much faster, and cheaply. One additional benefit was that the phone data could provide a much higher granularity in providing characteristics. Indeed, it was difficult to verify the quality of the insights provided by the phone data in Rwanda, because no other data source was able to provide wealth information with such geographical resolution.

<3.3> Refugees

Mobility data is particularly useful to study groups that are not easily covered in the national surveys, such as refugees. An example initiative is the Data for Refugees (D4R) Challenge, which was a data collaborative initiated by Turk Telekom, Boğaziçi University and TUBITAK and in collaboration with several academic and non governmental or intra-governmental organizations, including UNHCR Turkey, UNICEF, and IOM, to gain insights for improving the living conditions of over (then) three million Syrian refugees[ix] living in

Turkey (Salah et al. 2018; Salah et al. 2019). In D4R, an additional flag was used to tag each mobile CDR as possibly originating from a refugee, in case the phone line was 1) obtained with a subsidized, cheaper tariff for Syrian refugees, 2) registered with a Syrian passport, 3) registered with a special ID number that the Turkish government was providing the Syrian refugees. Each of these conditions contained some unspecified noise. Consequently, it was not possible to associate a CDR with a refugee with certainty. However, it was possible to derive conclusions from aggregated data on internal migration. For example, it was possible to see how many refugees were coming to the city of Ordu to work at the hazelnut harvest, which cities they were coming from, how long they were staying, and how they were distributed in and around Ordu during the harvest (Bruckschen et al., 2019; Turper Alisik et al., 2019). The dataset opened for the challenge contained data from 200K refugees and 800K Turkish citizens, and CDR were collected over one year (2017). 30 groups completed the challenge successfully, and submitted project reports[x].

As in all research involving mobile CDR, additional data sources or indicators must be combined with such data to investigate more complex aspects of mobility. For example, Beine et al. (2019) employed a gravity model to empirically estimate a series of determinants of refugee movements using the D4R challenge data, in order to evaluate how policy can facilitate mobility and integration of refugees. They considered standard determinants such as province characteristics, distances across provinces, levels of income, network effects, as well as refugee-specific determinants and the effect of certain categories of news events, protests, violence, and asylum grants.

<3.4> Humanitarian response

Humanitarian response requires up to date data processing for operational capabilities and informed decision making. Mobile CDR has been used by humanitarian organizations like UNHCR (Earney and Jimenez, 2019), UN Global Pulse (Boy et al., 2019), and UNICEF (Sekara et al. 2019) in refugee scenarios, but there are other crisis response settings, where its use have been demonstrated as well. One of the first applications was the analysis of people's movements in the aftermath of the 2010 earthquake in Haiti (Bengtsson et al., 2011). Mobility during floods, and the spread of epidemics, such as cholera, malaria, Ebola, and COVID-19, were also analysed and modeled with mobile CDR (Sandvik et al., 2014).

An important use of CDR is conducting pre-analysis to develop crisis response strategies and information preparedness. However, most research demonstrating the potential of mobile CDR analysis for humanitarian response is typically performed in the aftermath of emergencies, involving a significant delay. A major reason for this is that the sharing and processing of mobile CDR requires technical capabilities and resources, carefully prepared legal agreements, and ethical committee approvals that take time. However, researchers have discussed the possibility of preparing data processing pipelines and technical infrastructures well in advance of crises. A wake-up call was the COVID-19 pandemic, which will surely not be the last pandemic the world is facing. In (Oliver et al. 2020), we have stressed the importance of being ready for the next one, by creating data collaboratives that can be activated rapidly when a new epidemic is in its initial stages.

The need for actionable data is widely acknowledged, and there are important initiatives in this area. UN Global Pulse was established in 2009 to leverage big data sources in studying population behaviours for understanding and responding to global crises. A key role is played by coordinating agencies such as UN Office for the Coordination of Humanitarian Affairs

(OCHA), and by data holders (such as telecommunication operators) willing to act rapidly for humanitarian response.

<3.5> Processing mobile CDR

A detailed exposition of how mobile CDR data (or the higher-frequency x-Detail Record - xDR, which is obtained from data package exchanges between phones and operators) can be processed is beyond the scope of this chapter. Tools specifically developed for analysing mobility data include algorithmic packages like *scikit-mobility* (Pappalardo et al., 2019), and visualization tools such as *Urban Mobility Atlas*, which visually summarizes a number of mobility indicators over a geographical area (Gianotti et al., 2016). A good starting point that surveys advances made recently in the study of mobile phone datasets is (Blondel et al., 2015).

There are numerous factors and potential biases one needs to consider when analyzing mobile CDR data. Calabrese et al. (2013) list four of these as (1) the market share of the mobile phone operator from which the dataset is obtained; (2) the potential non-randomness of the mobile phone users; (3) calling plans which can limit the number of samples acquired at each hour or day; and (4) number of devices that each person carries. Additional factors include uneven gender distribution in phone ownership (e.g. in the D4R Challenge in Turkey, over 75% of phones were registered on male users, even though the proportion of females using the phones was probably much higher than 25%), and the lack of children, as they cannot legally own a phone line before a certain age. These are called 'data gaps', and they should be carefully integrated into the models of analysis. Zagheni and Weber (2015) propose calibration of data with reliable official statistics, when they are available, and evaluating relative trends, when such data are lacking.

<4> Ethics and privacy

Using big data technologies that provide granular and detailed view into human behaviour raises a number of important concerns that need to be addressed carefully. This is not only true of mobile phone data, but also of for example satellite imaging to detect human movements, or of using social media to chart out opinions and influences in social networks. These technologies, in the hands of controlling governments, can easily usher in an unprecedented degree of surveillance. As an example, in February 2020, the press reported that the U.S. Immigration and Customs Enforcement agency (ICE) has purchased commercial mobile phone datasets and used the information to arrest undocumented migrants (Tau and Hackman, 2020). Furthermore, irresponsible handling of data can harm the 'data subjects,' turning a social good application into a source of harm for the population it is intended to help (Berens et al., 2018). Subsequently, it is important to ask the question of what the risks are in designing technology and algorithms for such data analysis, and whether they are worth the potential benefits (Maxmen, 2019).

For each of the data collaboratives described in Section 2.3, a special committee was formed to screen proposals and reports, observing the 'Do no Harm' principle for the target populations (Vinck et al., 2019). In D4D, the D4D External Ethics Panel (DEEP) assessed applications for business ethics and intended applications, to balance risks and opportunities. This goes beyond what is permissible legally, and the panel was able to request amendments to reports or papers when necessary. Later, the Institute of Business Ethics (IBE) penned a report called 'Data for Development Senegal: Report of the External Review Panel' to share this framework and the findings of DEEP. The report stressed the importance of group privacy, as well as the importance of cultural factors, stating that "a particularly sensitive challenge is related to the lack of a regulatory framework or widely accepted benchmark on the privacy of data for individuals and for groups and its subsequent use or sharing. Another is the consideration for the publication of scientific results, the application of which might be considered more sensitive in certain cultures."

The D4R Challenge followed the recommendations of DEEP, and a project examination committee (PEC) was formed with members from academia, IGOs/NGOs, the telecommunications company initiating the challenge, representatives from two ministries of Turkey related with the challenge topics (i.e. education and health), as well as Syrian refugees. Together with the scientific committee, PEC evaluated all project proposals to determine who gets access to the data, and later evaluated all project reports and publications for potential risks. Once the challenge was over, PEC was disbanded, and it was not possible to screen later publications based on the D4R data.

The most important principle used in D4R was "data protection by design and default," where any name, real phone number, or other identifying information was excluded from the design of the database. The pseudo-random numbers representing customers were not stored anywhere along with actual phone numbers. Subsequently, the anonymization worked only one way. Since the refugee status was indicated by purposefully noisy indicators, and without any effort spent to ensure its validity, person-level conclusions about refugees could not be drawn from the data.

In all these challenges, the legal team of the data owner prepared an extensive license agreement for the challenge participants. Because the datasets did not contain personally identifying information, they were not considered personal information by definition, and did not legally require subscriber consent beyond what was specified in the mobile user agreement. However, the preparation of a legal document that maintains a high standard of accountability for data users is essential. To better understand the legal conditions that can enable effective data collaboration, GovLab, SDSN TReNDS, University of Washington's Information Risk Research Initiative, and the World Economic Forum have recently started an initiative called Contracts for Data Collaboration (C4DC)[xi], and created an online repository of data sharing agreements that have facilitated data sharing in a variety of settings and countries. These agreements cover a range of data applications and consider a host of legal issues that differ from country to country. The repository also contains associated use cases, guides and other resources.

The ICRC Data Protection Impact Assessment (DRIA) template[xii] provides a good framework that addresses many issues in mobile CDR processing in a systematic manner, providing a number of data protection issues together with related code of conduct, an example and practice oriented assessment of risks, potential mitigation measures, and potential outcomes for each issue, whereby a given scenario can be assessed. We provide a summary of the highlighted issues in Box 1. This, of course, is only one potential framework that can be used to address the issues that arise in using mobile CDR for migration and mobility. Several other frameworks are contrasted and detailed (see Berens et al., 2016).

- Purpose specification: The data should be collected only for a specific purpose.
- Data limitation: There should be no personal data items that are not required for analysis.
- Right to information: The individuals should be informed about collection and use of personal information.
- Legal basis for data processing / transfer: Informed consent should be obtained, under conditions where individuals giving consent are able to assess positive or negative outcomes.
- Right to access / rectification / deletion: Individuals should be able to access, correct, and delete their personal information.
- Information quality and accuracy: Processes should be in place to ensure information quality and accuracy.
- Appropriate security measures: The information should not jeopardize people, and must be appropriately protected against malicious uses, including surveillance.
- Data sharing, disclosure/publication and/or transfer: Data sharing with third parties and national societies should be transparent, meaningful, and accountable.
- Data retention: Data should not be retained longer than necessary.
- Risks to individuals: Additional risks, such as risks to physical or moral integrity of individuals, should be assessed.
- Accountability / oversight mechanism: Mechanisms should be in place to ensure responsible code of conduct and data protection standards.

Box 1: Key issues in ICRC Data Protection Impact Assessment (DRIA) template.

<5> Conclusions and key implications for research and governance

The effort in data challenges with mobile information illustrated that it takes a long time until research results can even begin to influence policy making. However, mobile CDR can potentially be processed very fast, much faster than traditional data acquisition mechanisms used by states. It is clear that the state would benefit from such faster decision making for informed policy decisions, but the potential issues of such processing are not yet tested, not even with the recent data challenges.

Many scientific papers were published on the findings of the large data challenges mentioned in this chapter, international media and NGOs showed great interest in the results, and several other initiatives were inspired. However, in many cases policy implications were limited, as the projects were exploratory. One proposed solution to increase impact for governance was to secure government funding for pilot projects following the challenge. Other strategies were to involve representatives from related ministries in the organization of the challenges, and preparing white papers for informing policy makers. What is still lacking at the time of writing this chapter is a data processing pipeline involving private data sources, designed in a privacy-aware way to inform authorities on urgent issues and help policy making (Letouze and Oliver, 2019; Oliver et al. 2020).

The value of mobile CDR for studying displaced populations is mainly in providing a rich data source that complements official statistics, and under certain safeguards, one that can give insights in real-time. This is particularly important for crisis management, where unexpected and large-scale displacement happens, or for studying challenging questions, such as the integration prospects of arriving migrants. However, the safeguards of using mobile phones as a data source without compromising privacy of individuals are just as important as the potential benefits.

It is clear that policy makers should continue an interdisciplinary dialogue to evaluate and discuss the positive and potentially negative consequences of mobile data processing. A clear understanding of how these technologies operate and what they can provide is essential for enabling the potential of using mobile data for social good.

Acknowledgments

We thank Almıla Akdağ, Marie McAuliffe, Bilgeçağ Aydoğdu, and Metehan Doyran for their constructive comments. This work is supported by the HumMingBird project, funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 870661.

References

Ahas, R and Ü. Mark (2005), 'Location Based Services—New Challenges for Planning and Public Administration?', *Futures*, **37**(6), 547–561.

Ahas, R., S. Silm and M. Tiru (2018), 'Measuring transnational migration with roaming datasets', paper in *Adjunct Proceedings of the 14th International Conference on Location based services*, ETH Zurich, 105-108.

Bakker, M.A., D.A. Piracha, P.J. Lu, K. Bejgo, M. Bahrami, Y. Leng, J. Balsa-Barreiro, J. Ricard, A.J. Morales, V.K. Singh and B. Bozkaya (2019), 'Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey', in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 123-140.

Bayir, M. A., M. Demirbas and N. Eagle (2009), 'Discovering spatiotemporal mobility profiles of cellphone users', paper presented at the *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops,* IEEE.

Beine, M., L. Bertinelli, R. Cömertpay, A. Litina, J.F. Maystadt and B. Zou (2019), 'Refugee Mobility: Evidence from Phone Data in Turkey', in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 433-449.

Bengtsson, L., X. Lu, A. Thorson, R. Garfield and J. Von Schreeb (2011), 'Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti' *PLoS Med*, **8**(8), e1001083.

Berens, J., U. Mans and S. Verhulst (2016), ‘Mapping and comparing responsible data approaches,’ available at *SSRN*, accessed 2 October 2019 at http://dx.doi.org/10.2139/ssrn.3141453.

Blondel, V.D., M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda and C. Ziemlicki (2012), ‘Data for development: the D4D challenge on mobile phone data,’ *arXiv preprint arXiv:1210.0137*.

Blondel, V.D., A. Decuyper and G. Krings (2015), 'A survey of results on mobile phone datasets analysis', *EPJ data science*, **4**(1), 10.

Blumenstock, J., G. Cadamuro and R. On (2015), 'Predicting poverty and wealth from mobile phone metadata', *Science*, ***350***(6264), 1073-1076.

Boy, J., D. Pastor-Escuredo, D. Macguire, R.M. Jimenez and M. Luengo-Oroz (2019), 'Towards an understanding of refugee segregation, isolation, homophily and ultimately integration in Turkey using call detail records' In A.A. Salah, A. Pentland, B. Lepri, E. Letouzé, (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 141-164.

Bruckschen, F., T. Koebe, M. Ludolph, M.F. Marino and T. Schmid (2019), 'Refugees in undeclared employment—a case study in Turkey', in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 329-346.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira Jr and C. Ratti (2013), 'Understanding individual mobility patterns from urban sensing data: A mobile phone trace example', *Transportation Research Part C: Emerging Technologies*, **26**, 301-313.

Campbell, S. W. and N. Kwak (2010), 'Mobile communication and civic life: Linking patterns of use to civic and political engagement', *Journal of communication*, **60**(3), 536-555.

de Montjoye, Y.-A., C.A. Hidalgo, M. Verleysen and V.D. Blondel, (2013), 'Unique in the crowd: The privacy bounds of human mobility', *Scientific Reports,* **3**:1376.

de Montjoye, Y.-A., Z. Smoreda, R. Trinquart, C. Ziemlicki and V.D. Blondel (2014), ‘D4D-Senegal: the second mobile phone data for development challenge,’ *arXiv preprint arXiv:1407.4885*.

de Montjoye, Y.A., S. Gambs, V. Blondel, G. Canright, N. De Cordes, S. Deletaille, K. Engø-Monsen, M. Garcia-Herranz, J. Kendall, C. Kerry and G. Krings (2018), 'On the privacy-conscientious use of mobile phone data', *Scientific data*, *5*.

Deville, P., C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel and A.J. Tatem (2014), ‘Dynamic population mapping using mobile phone data’, *Proceedings of the National Academy of Sciences,* **111**(45), 15888-15893.

Di Lorenzo, G., M. Sbodio, F. Calabrese, M. Berlingerio, F. Pinelli, and R. Nair (2015), 'AllAboard: Visual exploration of cellphone mobility data to optimise public transport', *IEEE Transactions on Visualization and Computer Graphics*, **22**(2), 1036-1050.

Dobra, A., N.E. Williams and N. Eagle (2015), 'Spatiotemporal detection of unusual human population behavior using mobile phone data', *PloS One*, **10**(3), e0120449.

Eagle, N. and A.S. Pentland (2006), 'Reality mining: sensing complex social systems', *Personal and Ubiquitous Computing*, **10**(4), 255-268.

Eagle, N. and A.S. Pentland (2009), 'Eigenbehaviors: Identifying structure in routine', *Behavioral Ecology and Sociobiology*, **63**(7), 1057-1066.

Earney, C., and R.M. Jimenez (2019), Pioneering Predictive Analytics for Decision-Making in Forced Displacement Contexts. In A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 101-119.

Giannotti, F., L. Gabrielli, D. Pedreschi and S. Rinzivillo (2016), 'Understanding human mobility with big data', in *Solving Large Scale Learning Tasks. Challenges and Algorithms,* Springer, Cham, 208-220.

González, M.C., C.A. Hidalgo and A.L. Barabasi (2008), 'Understanding individual human mobility patterns', *Nature*, **453**(7196), 779.

Gundogdu, D., O. Durmaz Incel, A.A. Salah and B. Lepri (2016), 'Countrywide arrhythmia: emergency event detection using mobile phone data', *EPJ Data Science*, **5**(1), 25.

Hughes, C., E. Zagheni, G.J. Abel, A. Sorichetta, A. Wi'sniowski, I. Weber and A.J. Tatem (2016), 'Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data: Feasibility study on Inferring (labour) mobility and migration in the European Union from big data and social media data', Report prepared for the European Commission project #VT/2014/093.

Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland and A. Varshavsky (2011), 'Identifying important places in people’s lives from cellular network data', in: K. Lyons, J. Hightower and E.M. Huang (eds) *Pervasive computing,* **6696**. Springer, Berlin, 133-151.

Isaacman, S., V. Frias-Martinez and E. Frias-Martinez (2018), 'Modeling human migration patterns during drought conditions in La Guajira, Colombia', paper presented at *1st ACM SIGCAS Conference on Computing and Sustainable Societies,* 1-9.

Jerven, M. (2014), 'Benefits and costs of the data for development targets for the post-2015 development agenda,' *Data for Development Assessment Paper*, Copenhagen Consensus Center.

Kang, C., Y. Liu, D. Guo and K. Qin (2015), 'A generalized radiation model for human mobility: spatial scale, searching direction and trip constraint', *PloS one*, **10**(11).

Lazer, D., A.S. Pentland, L. Adamic, S. Aral, A.L. Barabási, D. Brewer, D., ...and M. Van Alstyne (2009), 'Life in the network: the coming age of computational social science', *Science*, **323**(5915), 721-723.

Letouzé, E (2019), 'Leveraging Open Algorithms (OPAL) for the Safe, Ethical, and Scalable Use of Private Sector Data in Crisis Contexts,' in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 453-464.

Letouzé, E. and N. Oliver (2019), 'Sharing is Caring: Four Key Requirements for Sustainable Private Data Sharing and Use for Public Good', *DataPop Alliance - Vodafone Institute for Society and Communications White Paper.* Accessed: http://datapopalliance.org/wp-content/uploads/2019/11/DPA_VFI-SHARING-IS-CARING.pdf

Lu, X., D.J. Wrathall, P.R. Sundsøy, M. Nadiruzzaman, E. Wetter, A. Iqbal,... and L. Bengtsson (2016), '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*, **38**, 1-7.

Margolin, D., Y.R. Lin, D. Brewer, D. Lazer (2013), 'Matching data and interpretation: Towards a rosetta stone joining behavioral and survey data,' paper presented at the *Seventh International AAAI Conference on Weblogs and Social Media*.

Maxmen, A. (2019), 'Can tracking people through phone-call data improve lives?' *Nature News Feature,* 29 May 2019, Accessed: https://www.nature.com/articles/d41586-019-01679-5

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., Letouzé, E., Salah, A.A., Benjamins, R., Cattuto, C. and Colizza, V., (2020), 'Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle', *Science Advances,* **6**(23), eabc0764.

Palchykov, V., M. Mitrović, H.H. Jo, J. Saramäki and R.K. Pan (2014), 'Inferring human mobility using communication patterns', *Scientific reports*, **4**(1), 1-6.

Pappalardo, L., G. Barlacchi, F. Simini and R. Pellungrini (2019), 'scikit-mobility: An open-source Python library for human mobility analysis and simulation', *arXiv preprint arXiv:1907.07062.*

Reades, J., F. Calabrese, A. Sevtsuk and C. Ratti (2007), 'Cellular census: explorations in urban data collection', *IEEE Pervasive Computing,* **6**(3):30-38.

Salah, A.A. and T. Gevers (2011), *Computer analysis of human behavior*. London: Springer.

Salah, A.A., A. Pentland, B. Lepri, E. Letouzé, P. Vinck, Y.-A. de Montjoye, X. Dong and Ö. Dağdelen, (2018), ‘Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey,’ *arXiv preprint arXiv:1807.00523*.

Salah, A.A., A. Pentland, B. Lepri, E. Letouzé, Y.-A. de Montjoye, X. Dong, Ö. Dağdelen and P. Vinck (2019), ‘Introduction to the data for refugees challenge on mobility of Syrian refugees in Turkey,’ in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 3-27.

Sandvik, K.B., M.G. Jumbert, J. Karlsrud and M. Kaufmann (2014), 'Humanitarian technology: a critical research agenda,' *International Review of the Red Cross*, **96**(893), 219-242.

Šćepanović, S., I. Mishkovski, P. Hui, J.K. Nurminen, and A. Ylä-Jääski (2015), 'Mobile phone call data as a regional socio-economic proxy indicator,' *PloS one*, **10**(4).

Sekara, V., E. Omodei, L. Healy, J. Beise, C. Hansen, D. You,... and M. Garcia-Herranz (2019), 'Mobile Phone Data for Children on the Move: Challenges and Opportunities' in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 53-66.

Tau B. and M. Hackman (2020), 'Federal Agencies Use Cellphone Location Data for Immigration Enforcement,' *The Wall Street Journal*, Feb 7, 2020. Accessed: https://www.wsj.com/articles/federal-agencies-use-cellphone-location-data-for-immigration-enforcement-11581078600

Turper Alisik, S., D.B. Aksel, A.E. Yantac, I. Kayi, S. Salman, A. Icduygu, D. Cay, L. Baruh and I. Bensason (2019), 'Seasonal Labor Migration Among Syrian Refugees and Urban Deep Map for Integration in Turkey' in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 305-328.

Verhulst, S. G. and A. Young (2019), 'The potential and practice of data collaboratives for migration,' in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 465-476.

Vinck, P., P.N. Pham and A.A. Salah (2019), '"Do No Harm" in the Age of Big Data: Data, Ethics, and the Refugees', in A.A. Salah, A. Pentland, B. Lepri, E. Letouzé (eds.) *Guide to Mobile Data Analytics in Refugee Scenarios*, Cham, Switzerland: Springer Nature Switzerland AG, 87-99.

Wang, Y., L. Dong, Y. Liu, Z. Huang, and Y. Liu (2019), 'Migration patterns in China extracted from mobile positioning data', *Habitat International*, **86**, 71-80.

Zagheni, E. and I. Weber (2015), 'Demographic research with non-representative internet data', *International Journal of Manpower*, **36**(1), 13-25.

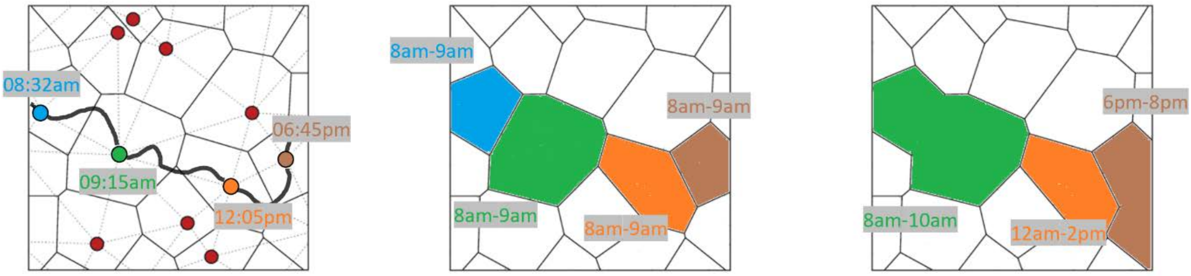

Figure 1: Anonymization and spatiotemporal aggregation of CDR data. (a) A user has four recorded calls in the telecom database, with precise call times, call duration, source and target base stations. The phone number is first replaced with a pseudorandom number. (b) The temporal granularity is decreased by replacing exact times with time intervals. The larger the interval, the coarser the granularity. Choosing one or two hours is typically enough. (c) Spatial granularity is decreased by joining adjacent cells. For tracks that are longer than a few weeks, such a step is a necessary precaution. Figure adapted from (de Montjoye et al., 2013).

[i] The International Telecommunications Union (ITU) estimates the percentage of the population covered by a mobile-cellular network to be 98,8% in the developed world and 96,2% in the developing world, according to 2019 data. Detailed statistics are available at: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

[ii] The anonymized Reality Mining dataset, collected over the course of nine months, can be downloaded from http://realitycommons.media.mit.edu/realitymining.html.

[iii] The GSMA Digital Toolkit includes detailed information on policy and regulations, business models, as well as a portfolio of case studies, and can be accessed at: https://aiforimpacttoolkit.gsma.com/.

[iv] See more on the SoBigData project at: http://project.sobigdata.eu/ and the HummingBird project at: https://cordis.europa.eu/project/rcn/225807/en

[v] Contributions to all past editions of NETMOB can be accessed at: https://netmob.org/#pasteditions.

[vi] The scientific reports of all D4D projects can be accessed at: https://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf.

[vii] Unlike the D4D and D4R Challenges, Telecom Italia Challenge made its data public for a limited time: http://theodi.fbk.eu/openbigdata/

[viii] It is important to note here that when the database is appropriately anonymized, such profiling can indicate a group membership with a high probability, but is not connected to a specific identity. Nonetheless, if sufficiently many indicators are added to the analysis, it becomes possible to pinpoint to an individual about which a lot of information is externally obtained. For this reason, both spatial and temporal resolution of the data are reduced on purpose.

[ix] Turkey is party to the 1951 Geneva Refugee Convention, but only acknowledges "refugee" status for people originating from Europe. Syrian refugees are officially and legally considered "temporarily protected foreign individuals".

[x] The scientific reports of all D4R projects can be accessed at: https://webspace.science.uu.nl/~salah006/d4r-proceedings.pdf.

[xi] Contracts for Data Collaboration (C4DC) website contains more information at: https://contractsfordatacollaboration.org/

[xii] ICRC Data Protection Impact Assessment template can be accessed at: https://www.icrc.org/en/download/file/18149/dpia-template.pdf.