12 June 2013

Review

Call Detail Records

The use of mobile phone data to track and predict population displacement in disasters

Contents

Acknowledgements

ACAPS wishes to express its gratitude to Nigel Woof for its time, dedication and enthusiasm in undertaking this research.

Abbreviations and acronyms

2G, 3G, 4G	Second (etc) generation of mobile phone networks and facilities. 2G networks offer calls and SMS messaging, but only limited data services.
CDR	Call Detail Record
GSM	Global System for Mobile Communications (originally Groupe Spécial Mobile)
BTS	Base Transceiving Station
BSC	Base Station Controller
DP	Displaced person (see also: IDP)
GIS	Geographical Information Systems
IASC	Inter Agency Standing Committee
IDP	Internally Displaced Person
MIRA	Multi Sector/Cluster Initial Rapid Assessment (a humanitarian needs assessment methodology currently being introduced)
MNO	Mobile Network Operator
MSC	Mobile Switching Centre
MVNO	Mobile Virtual Network Operator (a mobile phone company utilising another operator's actual network)
PSD	Preliminary Scenario Definition (the first situation analysis report in the MIRA approach to rapid needs assessment)
SIM	Subscriber Identity Module – usually termed a 'SIM card'.
XML	Extensible Markup Language (a data exchange format in which the file structure is pre-defined and fields are delimited using 'tags')

Executive summary

Information about the displacement of people after disasters is crucial in determining the scale and impact of the emergency, and is vital for conducting humanitarian needs assessment on the ground. Methods to forecast or detect such migration are however very limited at present.

The use of geo-referenced mobile phone call data to understand post-disaster movements of affected people has been demonstrated in two studies, in the aftermath of the Haiti (2010) and Christchurch, New Zealand (2011) earthquakes. These studies, matched against aid agencies' recurring information needs in disaster response operations, suggest that this type of data has potential to be a useful new method to forecast and locate people who have been displaced and therefore in need.

The Haiti and New Zealand studies showed that pre-disaster mobile phone usage patterns are highly predictive of where people will move when displaced by an emergency. Analysis of basic mobile call records (known as CDRs) is a practical method for inferring these migrations, to a useful degree of accuracy. The acceleration of mobile phone usage in developing countries should enable the practical use of call data in this way in many disaster incidents; however they have not yet been used as such since 2011. This is because, probably, the methods require substantial technical resources and, crucially, ready access to call data sets from mobile network operators: such cooperation is costly in time and effort and there are a number of institutional obstacles to be overcome, notably involving gaining access to the data, before data can be shared and used as envisaged.

While acquiring and analysing data from a 'standing start' in new disaster may remain problematic for the above reasons, the Haiti and New Zealand studies showed the potential of pre-analysing CDR data as a disaster preparedness activity, and then holding the analysed information products, cross referenced to other key baseline datasets for information preparedness. This approach would sidestep the logistical problems of attempting to capture, analyse and publish new datasets in the immediate aftermath of a disaster event.

Operationalizing these methods remains constrained by both availability of technical resources and also by non-technical factors including institutional, legal, cultural and commercial dimensions. While standardised common approaches have been advocated by multilateral actors, at present it appears that these issues will need to be addressed country by country for the time being.

1. Introduction

This paper describes the potential use of anonymised mobile phone activity data as a way to predict the movement of groups of people in disasters. It has been written ahead of practical trials of such methods, to be conducted by ACAPS with other partners during 2013. This paper summarises the current 'state of the art' of the methods and discusses their application. It draws on published research and also on interviews and discussions with researchers, practitioners and other stakeholders.

The number of people affected and/or displaced, and where they have moved to, is one of the most important factors to understand in any humanitarian emergency. Displaced people typically have the greatest and most urgent needs for relief assistance. Information about displacement can help to determine both the scale and the impact of a disaster. It will inform decisions about where to do on-the-ground assessments in order to initiate relief assistance. It is also the first step in establishing a monitoring system for the coordinated management of the response and recovery phases of the crisis. However, determining the extent and geographical picture of displacement, even approximately, remains difficult in humanitarian practice (ACAPS 2012).

Since 2006, locational patterns of mobile call use have been studied in social science research (Ajala 2006, Becker et al 2011, Isaacman 2011). Researchers have found that mobile phone usage gives good indications of the 'life patterns' (Tanahashi et al 2013) of populations, at least under non-emergency conditions. From that, it was speculated that the movement of people in disasters might conform to these 'life patterns', which have to do with social and family connectedness.

In the aftermath of the 2010 Haiti earthquake, Karolinska Institute with Columbia University and Digicel Haiti analysed anonymised data on movements of SIM cards from the Digicel Haiti mobile phone network, as a potential predictor of migration patterns following the earthquake. This yielded interesting results (Bengtsson et al 2011, Flowminder 2013): movement patterns pre-disaster, as shown by mobile phone data, tended to match closely post-disaster displacement outcomes. In 2011, following the Christchurch earthquake, Statistics New Zealand studied cellphone data records and analysed movement patterns of people to understand destinations of people who relocated from the city, and later their returns.

The Haiti and New Zealand studies showed how geo-located mobile calls records might in future be used to forecast likely displacement destinations, and possibly of the scale of displacement. This would give humanitarian actors an important new information resource to focus situational and needs assessment and to steer relief programming.

The backdrop to this paper is the growing attention paid to the applications of mobile technologies by humanitarian and development actors. Much research and development (see, for example, Vodafone Foundation and Save the Children 2013) has been focused on the proactive use of mobile phones for two-way messaging with affected people, and for cash-based relief programming. However, apart from the Haiti research referred to above, much less attention has been given to using the 'data exhaust'[1] of mobile phone networks as a humanitarian information resource. Recently, the potential applications of 'big data' as a humanitarian resource was explored in OCHA's report on *Humanitarianism in the Network Age* (OCHA 2013); the technical, methodological and institutional challenges of its use were also highlighted.

2. Predicting and tracking displacement

In most humanitarian emergencies it is displaced people (DPs)[2] who have the greatest needs, which are often urgent because they have left behind their basic resources and livelihoods. At the same time, an influx of DPs also puts pressure on host communities: a fact often insufficiently understood or accounted for in programming of response. Knowing the destinations of displaced people is therefore very important, both to assess the overall scale of the emergency and then to reach those displaced people, and those sheltering them, with relief assistance.

[1] The term 'data exhaust' refers to the data that is in effect a by-product of a digitally enabled activity; the user may not even be aware of its generation or collection.

[2] Displaced people may include both *internally displaced people* (IDPs) and *refugees*: that is people who have fled across an international border.

In major disasters or other humanitarian crises, detection, quantification and destination tracking of displaced people is typically extremely difficult. Even if it is known that people have moved away from their homes, there is a shortage of techniques to detect or forecast their destinations. Direct observation is unlikely to be effective and may not be feasible at all. Remote sensing (eg satellite imaging) has shown some promise in detecting camp locations and even movements of groups of people at a point-in-time; however it contributes at best only fragments of the overall picture of displacement over an entire emergency.

At present the destinations of DPs are most often determined through on-the-ground visits or surveys. However these typically take time to organise and – as a Catch-22 – it is obviously not possible to prioritise surveys on areas with high numbers of IDPs until destinations are known or can be reliably inferred. After the Haiti earthquake, the National Civil Protection Agency made estimates of displacement by counting numbers of buses and ships leaving the area (Bengtsson et al 2011). Rapid field assessments normally do not take the form of quantitative surveys, and while they may be good at characterising priority needs, they do not use statistically valid sampling[3] and are not expected to be effective at estimating numbers of cases, especially where displaced people are dispersed across a wide area. For example, in the conflict-driven crisis in Cote d'Ivoire in 2011, around 100,000 people are believed to have fled to camps and host communities over an area of more than 20,0000 square kilometres in the west of the country. However it proved impracticable to gain a usefully precise locational picture of the displaced people staying with host families, and most relief programming hence remained focused on the in-camps population[4].

Even if it is possible, on rare occasions, to survey accurately the numbers of people in new locations, accurate population baselines may not be available from which to infer where they have travelled from. This retains the problem of understanding the overall Humanitarian Profile[5] of the emergency.

It would therefore be very valuable if mobile phone data could contribute to the following:

- Forecasting of movements of population groups before they occur, by analysing past behaviours in both previous emergencies and non-emergency situations.
- Estimating the scale and destination of displacement at an early stage of a new emergency.
- Analysis of displaced people by area of origin, ethnicity and/or language, and economic status as implied by denomination of phone cruet purchases (assuming this data can be obtained from the data owner) as this may inform approaches to communicating with the displaced population and preparing to meet special needs.

This paper will focus on such applications in natural disasters. Situation analysis in 'complex' (that is, conflict-driven) humanitarian emergencies is particularly challenging. Some of the methods described may turn out to be applicable in complex crises also but this will need further study.

3. Opportunities, resources and challenges

Mobile phones in the developing world

The high take-up of mobile phone use globally, including particularly in developing countries, is a well-documented phenomenon and will not be explored in depth in this paper. Global mobile phone subscriptions have now exceeded 6 billion (Vital Wave 2013) and the strongest growth in take-up continues to be in the developing world (ITU 2012).

Relying on subscriber statistics to assess actual phone usage does however have some pitfalls. The growth in pre-paid phone accounts, which is driving market expansion in the developing world, means that some people have multiple phones, or at least multiple SIM cards, to enable them to exploit pricing and coverage variations; for example in Mozambique 65% of regular mobile phone users had access to more than one phone or SIM card (InterMedia 2010). The predominance of pre-paid accounts may also prove useful as an indicator of

[3] In rapid assessments it is usual to adopt *convenience* or *purposive sampling*, which implies only moderate or low external validity and hence poor reliability to support crucial decision making.

[4] Population displacement estimates by OCHA and MapAction, from joint field assessments.

[5] An IASC standard model for quantifying the humanitarian caseload by relevant population segments.

relative wealth, if the poorest users tend to buy very small top-ups: however this correlation has not been explored.

Further relevant characteristics of the Mozambique market in 2010, also from InterMedia and used here to typify many poorer countries' mobile phone markets, were as follows:

- Although Mozambique has lower mobile penetration than many other African countries, about one third of the adult population had mobile phones in 2010.
- Regular mobile phone users are skewed towards younger adults. 36% are 15 to 24; 29% are 25 to 34; and 30% are 35 to 50. However, only 5% of mobile users are above the age of 50, despite this age group being 17% of the adult population. Taken with the overall phone usage percentage in the population, it implies that only perhaps one in ten older people are regular phone users.
- 88% of users had completed at least some secondary education: well above the national average.
- Only 43% of users are female, and fewer still in rural areas.

Call records: the 'data exhaust' from mobile phone networks

Mobile (cellular) phone systems are now a mature technology. Older analogue systems have now been generally replaced by digital ones, almost all currently using the GSM standard for 2G and evolutionary standards for 3G and, soon, 4G. The majority of mobile phone networks therefore share similar infrastructure and operating systems. An overview of the relevant architecture of mobile phone networks is given in Appendix A.

Every mobile phone network captures and uses information about calls made by its individual users, primarily so that they can be billed (or a prepayment account can be debited). While the content of the voice call or SMS will not normally be recorded, the basic facts about the call (sometimes termed the 'metadata') are routinely captured under GSM standards in the form of Call Detail Records (CDRs). These records include the calling and called connections, time and duration of calls, use of various services (eg SMS), and the location IDs of the cells in use. CDR records are not entirely standardised and will vary depending on the operator, their systems software supplier, and the system configuration. The structure of this data is described in more detail in Appendix A.

It is worth noting that a network actually logs a very large amount of data about technical performance, of which a small proportion is the CDR. Some of this non-call data may well be valuable for humanitarian purposes; however at this stage we are considering only call data, in the form of CDRs. Most research studies to date, including the Haiti and New Zealand studies, have used batches of CDRs, obtained from mobile network operators, as their principal data source.

Mobile phone records as 'Big Data'

The information revolution is creating massively expanding amounts of potentially useful data, often as a by-product of business, public administration and consumer activities. The total volume of such data is reported to be doubling every 20 months (Global Pulse 2012). A decade or more ago, 'data mining' described the process of analysing large data sets using established statistical and database interrogation methods. However, the notion of 'Big Data' involves datasets that are so large and complex that they need a new generation of tools and methods to exploit them as a resource. Some of these analytics tools are designed to be inherently 'sense making' as they embody algorithms that search for patterns in the data.

The potential of Big Data in the humanitarian sphere has not yet been explored in any real depth. This is in large part because developing countries are only starting to see the creation of Big Data as digital activity expands in consumer, business and public administration spheres. However mobile phone usage is an exception: it is not constrained by limited broadband internet connectivity in developing countries (indeed, mobile phones are in many ways a substitute for it) and the capture of mobile phone activity data is equally feasible whether the network operator is in a rich or poor country.

Given also the continued strong interest in the exploitation of mobile phone technologies for social benefit in developing countries, it is not surprising that there is growing interest in the potential of call data. This does mean however that mobile phone companies are receiving growing numbers of requests for data for a wide range of such purposes.

Exploiting Big Data for public benefit is increasingly seen as involving combinations of data from multiple sources (Boyle 2013). A paper presented at the World Economic Forum (Vital Wave 2013) argued for a 'data commons' approach and evolution of international norms and a code of conduct to lower barriers to productive sharing of anonymised data. However, such a harmonised approach may not be imminent.

The notion of Big Data is sometimes conflated or even confused with that of Open Data, particularly in development and humanitarian contexts. However, there are important differences between these concepts. Most importantly for the purposes of this paper, Big Data may not be fully open, but may instead remain proprietary (Global Pulse 2012): this is especially likely to be the case with mobile phone data, at least in the near term.

Access to mobile phone call datasets

CDR data is the property of the mobile network operator and has a high commercial value, particularly to competing operators. The data also has value to other types of business: for example to a retail chain undertaking store location analysis, or for advertising planning.

Motivations of operators to release data either before or during an emergency might be driven by corporate responsibility considerations. Sources at the GSM Association stress the importance of a clear business case for the positive effort necessary to enable data release. From the operator's perspective, any data release involves apparent effort and risk, with less clear benefits. There may be contractual constraints on operators over release of customer data, or at least a perceived loss of consumer trust if records are transferred outside the company's control.

Privacy is important not only to maintain public acceptance of data usage, but because data might be used deliberately against vulnerable individuals or groups, compromising their human rights. On the other hand it may be that governments wishing to access and abuse such data will in any case obtain data directly from mobile operators, overtly or covertly: many countries have 'intercept laws' which, as they stand or in a subverted form, may enable this to happen. Methods of achieving anonymisation to preserve privacy are descried briefly in Appendix B.

4. What was learnt from the Haiti and New Zealand studies

Haiti

The Karolinska/Columbia work with mobile phone data in Haiti in 2010 was referred to in the introduction to this paper and documented in a series of four research output reports (Bengtsson et al 2010-11) and several peer-reviewed papers (including Bengtsson et al 2011, and Lu et al 2012). It is the most important research resource for the purposes of this paper. It was a highly innovative and successful attempt to acquire, analyse and use CDR data to triangulate against other information about population displacement after a disaster – the 12 January 2010 earthquake. The team also attempted to apply the same methods in near-real time during the cholera outbreak in October the same year.

The following is summary of key findings from both parts of the study. The most useful comprehensive report on the work is contained in the Bengtsson et al (2011) paper.

1. Post-earthquake migration analysis

The study team analysed CDRs from 1.9 million Digicel phones to track people moving out of (and later into) Port-au-Prince (PaP) metro area. The first analysis was released in May, four months post-disaster. It identified numbers and timing of out-migration from PaP to all Haitian departments (districts), which peaked on EQ+19 days. The analysis showed clearly that people took longest to reach the most distant departments. The team then analysed later months' data and graphed the gradual return to PaP over the next ten months.

Phone usage pre-EQ, particularly over the Christmas/New Year period, was analysed and compared with the post-EQ results. This showed a high correlation of post-EQ migration with the places people had visited over the holiday period. This suggests that displacement destination was aligned with the localities of social/family contacts, and hence that displacement destinations could have been predicted by reference to baseline CDR datasets.

The study team did not publish any analysis of migration destinations within the departments. It would however presumably have been technically possible to discern more precisely the destinations of displaced people, for example whether to major towns or rural communities, or indeed to camps, given the resources to undertake that level of analysis.

2. Study of cholera outbreak

A cholera outbreak was confirmed in the Saint-Marc district on 21 October. The study team responded quickly and analysed CDR data from the central departments, covering the week in which the outbreak was confirmed with a three-month period earlier in the year as a baseline for comparison. It is presumed that these datasets had already been acquired by the study team as part of the main study, making a fast analysis and issue of results possible.

The results allowed a statistically sound comparison of movements day by day around the outbreak date. In fact, the team reported that movement did not seem to be out of the ordinary, and conformed to baseline weekly commuting patterns. However, the team postulated that the destinations of mobile phone users who had moved out of the affected area (whether routinely or in flight) might be predictive of areas at risk of disease transmission.

Both stages of the research used only the location from which calls were made, without attempting to capture and analyse the destination location of the call. This undoubtedly simplified the research while illustrating the potential of the most basic level of data. It does leave open the question of whether analysing call destinations might add usefully to the results of future studies and in practice.

It was reported that the user populations of the two Haitian GSM mobile phone operators did not appear to differ significantly in profile. This justified the generalisation from only one company's data (Digicel). It might be more usual to find competing networks' subscriber bases differing in important characteristics, and this must be borne in mind in transferring the methods to other countries.

It is important to note that the two stages of the Haiti study employed different temporal modes in respect of the data used and the timing of processing and use. The post-earthquake analysis actually used pre-disaster data. It could be classified as *pre-disaster data with post-disaster analysis*. The cholera analysis on the other hand included analysis of both pre-outbreak (baseline) and new, post-outbreak data, obviously analysed post-outbreak: in shorthand perhaps *pre+post-disaster data with post-disaster analysis*. The relevance of this nomenclature for future practice will be discussed later in this paper.

New Zealand

After the 22 February 2011 earthquake in Christchurch, the NZ Ministry of Civil Defence and Emergency Management approached Statistics NZ, a government agency, for assistance in a number of areas. This led to a study of the potential of cellphone records to track population movements, with the results published (Statistics NZ, 2012).

The NZ study team acquired CDR data from one mobile operator, filtering the records to include only those that initiated calls from cellphone towers in Christchurch city, during the 10 days before the earthquake. They analysed movement of those phones for a period of more than a year before the disaster, and for several months after. As in the Haiti studies, this enabled a comparison of 'normal' patterns of movement, against post disaster movements: so the NZ study was *pre+post-disaster data with post-disaster analysis*.

The analysis gives clear indications of people's movements both pre- and post-disaster. Overall movement by the sampled phones outside the Christchurch region more than doubled in the six weeks post disaster, compared to the same period a year earlier. As in Haiti there was a strong correlation between movement destinations during the 'normal' and post-disaster periods. And, as in Haiti, migration did not spread outwards uniformly from the affected city. For, example many people travelled to Auckland, at the northern end of the country; while this might be explained by Auckland being the largest city in New Zealand, it also happened that large numbers of people relocated to low populated areas of the South Island (notably, Otago). The study's authors observed that established social connections were likely to be a factor in those cases. In fact, the study also detected that movements to outlying destinations took place later than closer ones: again, this was in some respects similar to Haiti.

The New Zealand study also examined ratios of voice calls and SMS messages: voice calls being normally about one third of all calls but more than half of all calls post-earthquake. The report did not find any significant differences between voice and SMS calls in terms of location tracking, and concluded that analysing voice calls only gives adequate results. This is important because, as the NZ team noted, some network operators may not retain CDRs for SMS calls.

The NZ study report offered a number of caveats on the methods, noting that cellphone data as sampled cannot detect migration for particular residential communities (it being impracticable to discriminate between business and non-business calling). Also, permanence of relocation cannot be determined. The team rightly concluded that actual numbers of people relocating cannot be detected from the CDR data; however, as with the Haiti study, it seems feasible that sampling could enable approximate estimates of the scale of movement of populations, to a useful degree.

5. Applications and stakeholders

The role of CDR-derived products in humanitarian needs assessment processes

While CDR-derived locational data has apparent potential to support humanitarian goals in numerous ways (for example, epidemiology), this paper is focusing on the development of information products that support situational analysis during the early response phase of major natural disaster emergencies, and particularly to predict population displacement for the purposes of humanitarian needs assessment. It may be that the methods under consideration, once properly tested, might turn out to be applicable in assessment of conflict-driven crises, but this needs thorough additional research focusing in complex emergency situations.

Recently the common approach to needs assessment has been re-visited by the international humanitarian community, through the IASC Needs Assessment Task Force. From this, there are some recognised principles of coordinated needs assessments that should inform the design of CDR-derived data initiatives for understanding migration during a crisis:

1. Knowledge about the pre-disaster context is very importance in inferring post-disaster scenarios (even though such knowledge has often been used poorly in practice by the aid community in new emergencies). Review and synthesis of baseline information, from a range of secondary sources, is now advocated as an important disaster preparedness practice – even though in some cases it is actually still done hurriedly at the onset of a new crisis (ACAPS 2012).

2. Knowledge of what happened in previous crises in the same country or area can give very valuable clues to impacts (including displacement) in a new emergency.

3. Once an emergency has begun, there is invariably a need to understand quickly the general characteristics of the situation and probable needs. The IASC Operational Guidance on Coordinated Assessments (IASC 2012) prescribes the Multi Secor/Cluster Initial Rapid Assessment (MIRA) approach in which an early situation analysis is done through the compilation of a Preliminary Scenario Definition (PSD), within 72 hours of a sudden-onset crisis. The PSD contains an estimation of the scale and severity of the event (‘how big’) and of its impact of the event (‘how bad’). The numbers and locations of displaced people, if known, provide key inputs to this analysis.

The need for a preparedness-based approach

Timeliness of CDR-derived information products is therefore of the highest importance if they are to be useful in supporting decision making. The ’72 hour window’ referred to above poses a considerable challenge if such information products are produced from a ‘standing start’ at disaster onset. This implies that a preparedness-based approach will be necessary.

Although the Haiti study was kicked off post-disaster, it is clear that there would be considerable value in having this class of predictive information on likely migration destinations available from the very onset of a new emergency. Because the CDR-driven approach involves using pre-crisis behaviour (and pre-existing locations of family and social contacts) as a predictor of where people will travel when a crisis occurs, there is obvious potential for the whole information resource to be addressed as

a disaster preparedness measure: in the proposed nomenclature, *pre-disaster data with pre-disaster analysis*. This follows the principle, exemplified by the IASC's COD/FOD standards[6], of information readiness for sudden onset disasters. While such an approach would presumably be focused on vulnerable countries and sub-territories, it does imply considerable resources to create the standby information products, while the perishability of such data is not known. It may therefore be a 'stage two' programme, once the more basic measures proposed in this paper have been proven.

Also, in some situations the use of post-disaster data would be precluded by any major disruption to the mobile phone network: if calls cannot be made, CDR records will not exist. In practice, mobile phone networks appear to be highly resilient (and increasing this further is a key aim of the industry, through its disaster preparedness initiatives coordinated through the GSM Association). Still, even if the network remains operational, ability to make calls will depend on the availability of electricity for recharging handsets. Moreover, placing demands on mobile phone providers for data at the crucial stage of a new emergency would obviously be less than ideal, again strengthening the case for a standby data approach, at least to create usable baselines.

Where it is not possible to create and maintain pre-processed datasets, and processing and analysis must therefore be done post-disaster, it would be realistic to accept significant limitations in reliability of analysis conclusions to meet decision time frames, provided that best efforts are made to assess the likely level of reliability and to caveat the use of the information accordingly. As with most information in humanitarian situations, a reasonably reliable answer now is almost always better than waiting for more thorough validation. Even a rough analysis from a preliminary review of the data may yield usable outputs, especially if they can be triangulated against other information. This follows the practical approach to needs assessment advocated by ACAPS, where "It is more important for emergency coordinated assessments to be rapid than it is for them to be detailed" (ACAPS 2012)

[6] Common and Fundamental Operational Datasets: a defined suite of baseline and other datasets that should be readily available at the onset of a new emergency.

Stakeholders

The applications explored in this paper necessarily involve a broad spectrum of stakeholders, to provide, process, and use the results from the process. The table below attempts to list the main groups of actors, both within the disaster-effected country and at an international level.

Table 1: Stakeholder groups

Category of actor	Data provider	Service provider	User of outputs
National actors			
Mobile phone users	Yes		Beneficial
Mobile phone network operators (companies)	Yes		
Government and regulatory authorities	Influence		
Academic and research institutions in affected country		Possibly	
National NGOs			Yes
International actors			
GSM Association	Influence		
OCHA			Yes
Other UN agencies and projects			Yes
International NGOs			Yes
International academic and research institutions		Possibly	
'Crisis mapping' community		Possibly	

The above list is unlikely to be exhaustive and other permutations of actors and roles are quite possible. Some of these stakeholders are required to

comprise the process chain to create the envisaged information products (possibly at multiple stages), while others are prospective users of the products, or may influence availability or use. Some may fulfil multiple roles. The characteristics and potential role(s) of some key actors are considered further below.

Mobile phone users. The methods involved in this paper involve the use of anonymised call records, and mobile phone users will presumably not be aware of the use of their data. However, they are obviously, as a community, the ultimate beneficial user of the information products provided to disaster preparedness and response actors. Good practice implies that consideration should be given to the application of the information in communicating with disaster affected communities (CDAC) initiatives, although specific applications in that context are outside the scope of this paper.

Mobile phone operators. Mobile network operators (MNOs) will be the primary data providers and hence crucial in all projects.

Operators range from large integrated global businesses such as China Mobile, Vodafone and Airtel that have subscribers numbered in hundreds of millions across multiple countries, to independent companies operating in a single national market. However, even small countries tend to have multiple competing operators except in a small minority of countries (eg Ethiopia) where the government telecoms concern holds a monopoly on mobile phone services. The GSM Association (GSMA), headquartered in London, facilitates cooperation between 827 mobile operating companies worldwide that use the GSM standard.

In many countries there are more sellers of mobile phone services than there are actual networks. Mobile Virtual Network Operators (MVNOs) may sell services under their own brand, while using the physical network of another operator. Ownership and control of CDR data where an MVNO is involved may vary.

Competitors often target different market segments within a country: for example one may focus on commercial sector users. Also, network geography may vary between operators which will affect takeup and usage. This means that market share figures based on subscriber registrations may not reflect the relative usage of competing networks by people most vulnerable in disasters – typically the poorest communities.

Regulators. In each country there will usually be one or more entities whose responsibilities include regulating the mobile phone market, including allocating frequency bands, policing relevant laws, ensuring fair competition, etc. Regulators may be government departments, or independent or quasi-independent. Regulators may have an influence on release of CDR data, either through explicit controls or through implicit regulatory concerns on the part of the subject MNO.

Data release and use may also be governed by data protection laws and regulators. In developing countries however, such frameworks are often nascent, or do not exist at all (Privacy International, 2012).

6. Practicalities

Process chains

As noted in the introduction to this paper, although the Haiti and New Zealand studies provide invaluable insights into the potential of the CDR data-driven methods, they have not yet been used to inform decision making in an emergency. Interviews and discussions in the preparation of this paper suggest some reasons for this slow exploitation of the apparent opportunity. In large part it is probably due to the complex processes and to the unique mix of stakeholders that it would be necessary to engage on each occasion.

For each application instance, the following process segments would appear to be required, as a minimum.

a. Understanding needs of humanitarian decision makers.
b. Project design, management and stewardship.
c. Access to appropriate CDR datasets, within an appropriate time frame.
d. Technical capacity for processing and analysis.
e. Processes to disseminate outputs through trusted humanitarian information channels (including the requirement to ensure protection from harm of data subjects).
f. Feedback and learning.

These processes will be explored during the practical trials to be undertaken by ACAPS and partners. Appendix B of this paper contains some of what is already known or envisaged as being involved in accessing and processing the data.

Information product design

In new crises, discovery of information resources is often a challenge, as humanitarian actors can be overwhelmed by data (UN Foundation 2011, OCHA 2013). They may have insufficient time even to consider using some potentially crucial information resources due to lack of awareness of their content or relevance. In addition to being relevant and actionable, this means that CDR-derived information must be made available in ways appropriate to the habitual processes being used for information acquisition by key actors. The outputs must be made available in formats that enable the intelligence they contain to be used in triangulation with other information sources.

As has been increasingly noted in recent years (eg in OCHA 2013), the supply of data to humanitarian actors following a major disaster typically reaches overload almost immediately. It is therefore essential that CDR-derived analyses are relevant and actionable, and moreover that they are aligned with meeting the critical situational intelligence requirements at key decision points. This implies that the information products as delivered need to have been designed by partners who understand well this user need.

The format by which humanitarian actors would wish to receive CDR-derived migration information is likely to vary with phases of the emergency and, to a lesser extent, with the aid organisation's sectoral focus and tasks. However it is very important that the format of the information is matched to the decision need.

Humanitarian users might require the following outputs from CDR-derived data analysis:

- Analysis reports in document form, including updates and any revised analyses.
- Summarised datasets, including data in GIS-ready form.

The design factors for these products, both in report form and as GIS datasets, are explored in more detail in Appendix C.

Another problem encountered in recent large disaster emergencies has been the re-dissemination of the same data resources by multiple organisations or even by individuals in the online crisis response communities, in slightly different formats but often with little value-added. This can lead to wasted effort on the part of users in reviewing multiple versions of essentially the same data, or confusion about data provenance. It would be beneficial if this tendency could be avoided with CDR-derived information. In any case, the provision of clear metadata[7] is very important, particularly the date(s) to which the data corresponds – which, as with many classes of emergency data, can often be confused with dates of analysis or dates on which the information product was issued.

It is also important that users are made readily aware of the existence of updated information products, to minimise the inadvertent use of older versions in circulation. One way to achieve this would be to maintain a single web page portal for all CDR reports and datasets with clear version referencing. Key metadata such as versions and date ranges should ideally also be included in filenames.

Issues and risks

The use of CDR-derived data in the ways envisaged creates some risks to the rights and welfare of people the methods are intended to help, and as with all humanitarian actions it is essential that these risks are considered and mitigated ethically. In compiling this paper, two types of risk have become obvious: (a) risks to individual data subjects (or to their families and communities) arising from their use of data about them; and (b) risks arising from erroneous use of CDR-derived information products by humanitarian actors, leading to false conclusions about needs that lead in turn to the wrong decisions being taken by humanitarian actors.

The issues of privacy and control of release of data were referred to earlier in this paper and are likely to be the concern of data owners, regulators and

[7] Metadata means 'data about data' and is an important consideration in humanitarian information management. In this context it means that CDR-derived datasets should be accompanied by notes about their provenance and other reference information.

other stakeholders, whether or not the data subjects themselves (mobile phone uses) are even aware of the collection and use of the call records data. The right to privacy has been argued to constitute an intrinsic right (eg, in Global Pulse 2012, p24) and this implies that it should be given strong weight even when it must be balanced against discharging the humanitarian imperative in an emergency. The New Zealand study (Statistics NZ, 2012) was able to be conducted in compliance the specific national privacy legislation and this may provide a useful starting-point model of data protection compliance for other jurisdictions. While tis paper considers natural disasters, the risks of inappropriate data disclosure might of course be even more serious in conflict-related crises and this would require extreme care.

In an emergency situation where there is actual or potential conflict or targeting of particular groups, the issue of protection (using the humanitarian sense of the term) becomes of vital importance. This will have implications for the handling and processing of data and about the dissemination of information products even where individuals or families cannot be discerned but where targeting of particular communities by armed groups is a possibility: displaced people are often particularly vulnerable in that respect. Ethical considerations about data and information release may also need to take into account situations where governments are suspected to infringe the human rights of displaced people, for example by forced resettlement or other coercive policies.

In the second category of risk, that of drawing incorrect conclusions from CDR-derived information products, this type of concern should always be addressed carefully in humanitarian practice but in the knowledge that a perfect or even highly reliable alternative information source is unlikely to exist but operational or resource allocation decisions still have to be made. Potential sources of bias in CDR-derived data have been highlighted in much of the published research referenced in this paper. The chief concern is that mobile phone ownership and use is not uniform across populations. The unequal use of technology within communities raises particular issues that were thrown into sharp focus by the consequences of the 2011 Japan earthquake and tsunami, in which older people suffered disproportionate numbers of deaths because they were not habitual mobile phone users and so did not receive tsunami warnings (Internews 2013). As the Haiti study team noted, mobile phone use may also be low among children, women, and the poorest.

Opportunities to make calls (and so generate CDRs) may also be biased by temporal or spatial variations in network coverage. Obviously, people moving to areas outside cell coverage, whether pre- or post-disaster will constitute gaps in the data.

7. Conclusions and next steps

The strong results from the Haiti and New Zealand studies, underpinned by a sound body of social science research in other contexts, suggest that CDR-inferred ‘spatial patterns of life’ have high potential in predicting population displacement in disasters. This is of great potential value, as it can fill a gap in the toolkit of available methods to determine rapidly the likely geographical extent and scale of a disaster ‘footprint’ (beyond the physical impact of the disaster), and ultimately to bring the most appropriate relief assistance to displaced populations in need. It remains to be shown whether the method can also detect numbers (scale) of displacement, possibly by comparison with other types of data.

Timeliness of CDR-derived information products is critical if it is to be useful during the crucial early stages of the MIRA assessment process: the target being 72 hours post-disaster. To acquire, process and analyse new, post-disaster, mobile phone data in the immediate aftermath of a disaster event is likely to pose considerable organisational and logistical challenges. However, the Haiti and New Zealand studies show the value of pre-disaster data, captured and processed during ‘normal times’ which can give valuable forecasts of post-disaster displacement behaviour. The ‘preparedness’ approach is probably more effective than attempting to commence CDR data acquisition from a standing start in a new emergency.

It would be ideal if the key stakeholder communities – including the mobile phone industry, regulators, and the humanitarian sector – could agree a common approach to the use of ‘big data’ for this purpose. Such approaches are indeed being advocated, for example by UN Global Pulse. This would, presumably, be a necessary precondition to the establishment of a ‘clearing house’ service for CDR data for disaster, as envisaged by the Flowminder project. However it may be some time

before such approaches are likely to gain traction. Meanwhile, the proposed CDR data application will probably need to be developed on a country by country basis.

For each national case, it will be necessary to develop a model for acquiring, processing, analysing and disseminating relevant CDR data to provide rapid forecasts and indications of displacement during the first 72 hours of a new sudden-onset disaster event. Due to the unique stakeholder mix for each country, differing technical standards and disaster hazards and baseline conditions, the approach will necessarily vary. However, there are bound to be common approaches that allow a semi-standardised methodology to be conceived. That will be a key aim for the proposed further trials by ACAPS and others. The design of these trials are outside the scope of this paper. However this paper highlights some factors affecting selection of countries for testing: most importantly there needs to be one or more mobile network operators who are willing and able to commit time and resources to the trial, and it would be ideal if patterns of previous natural disasters inform the formats for information products in a real-world context.

It is apparent that the processing of CDR records requires some expertise in handling and analysing large volumes of data. While there are international institutions with such capability, it is obviously important for national capacities to be developed as early as possible. It may in any case be necessary to use processing partners within the country concerned, in order to secure the cooperation of data owners and regulators.

It will also be important to involve humanitarian sector stakeholders in trials, to gain experience in designing CDR-derived information products that mesh with common needs assessment protocols, notably the MIRA.

Thanks are offered to all those who provided input to the paper and who kindly reviewed and commented at the drafting stage.

References

ACAPS (2012) **Coordinated Assessments in Emergencies: what we know now. Key lessons from field experiences**. Accessed from: http://awww.acaps.org/en/resources

Ajala I (2006): **Spatial Analysis of GSM Subscriber Call Data Records**. Directions Magazine, 7 March 2006.

Becker R A, Caceres R, Hanson K, Meng Loh J, Urbanek S, Varshavsky A, Volinsky C, (2011): **Clustering anonymized call detail records to find usage groups**. 1st Workshop on Pervasive Urban Applications (PURBA), 2011. Accessed from: http://www.research.att.com/people/Becker_Richard_A?fbid=CNwGe-6tG-p

Bengtsson L, Lu X, Garfield R, Thorson A, von Shreeb J (2010-11): **Internal population displacement in Haiti: preliminary analyses of movement patterns of Digicel mobile phones** (four reports). Karolinska Institute and Columbia University.

Bengtsson, et al, 2011: **Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti.** PLOS Medicine. Accessed from: http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001083.

Boyle P (2013): **Measured, recorded and then what? Lecture on Big Data applications for public policy research**. Given at Royal Geographical Society, London, by Paul Boyle, chief executive of UK Economic and Social Research Council, 4 March 2013. (Unpublished).

Flowminder (2013): **Flowminder project website**. http://www.flowminder.org/research/

Global Pulse (2012): **Big Data for development: challenges and opportunities**. Accessed from: http://www.unglobalpulse.org/projects/BigDataforDevelopment

IASC (2012): **Inter Agency Standing Committee Task Force on Needs Assessment**. Website: http://www.humanitarianinfo.org/iasc/pageloader.aspx?page=content-subsidi-common-default&sb=75

Intermedia (2010): **Mozambique: the who and what of the mobile phone market**. Accessed from: http://www.audiencescapes.org/country-profiles/mozambique-who-and-what-mobile-phone-market-mozambique-mcel-vodacom-SMS-tariff-profile

Internews (2013): **Connecting the last mile: report on the role of communications in the Great East Japan Earthquake**. Accessed from: http://www.internews.eu/News/Japanreport/

Isaacman S, Becker R, Caceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011): **Identifying important places in people's lives from cellular network data**. Accessed from: http://www2.research.att.com/~varshavsky/papers/isaacman11places.pdf

ITU (2012): **ITU releases latest global technology figures** (press release). Accessed from: http://www.itu.int/net/pressoffice/press_releases/2012/70.aspx#.UUBqSBzIYZk

Lu X, Bengtsson L and Holme P (2012): **Predictability of population displacement after the 2010 Haiti earthquake**. Proceedings of the National Academy of Sciences of the USA. Accessed from: http://www.pnas.org/content/early/2012/06/11/1203882109.abstract

OCHA (2013): **Humanitarianism in the Network Age**. Accessed from: http://www.unocha.org/node/11528

Privacy International (2012). **Privacy in the developing world: a global research agenda.** Accessed from: https://www.privacyinternational.org/blog/privacy-in-the-developing-world-a-global-research-agenda

Statistics New Zealand (2012): **Using cellphone data to measure population movements**. Accessed from: http://www.stats.govt.nz/tools_and_services/services/earthquake-info-portal/using-cellphone-data-report.aspx

Tanahashi Y, Rowland J R, North S, Ma K-L (2013): **Inferring human mobility patterns from anonymized mobile communication usage**.

Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia. Accessed from: http://vis.cs.ucdavis.edu/~tanahashi/PUBLICATIONS/YZR_MoMM2012.pdf

UN Foundation (2011): **Disaster Relief 2.0: the future of information sharing in humanitarian emergencies**. Accessed from: http://www.unfoundation.org/news-and-media/publications-and-speeches/disaster-relief-2-report.html

Vital Wave (2013): **Paving the path to a Big Data commons**. Briefing for World Economic Forum. Accessed from: http://www.vitalwaveconsulting.com/Insights/articles/2013/Big-Data.htm

Vodafone Foundation and Save the Children (2013): **Mobile Technology in Emergencies**. Accessed from: http://www.savethechildren.org.uk/resources/online-library/mobile-technology-emergencies

Zang H and Bolot J (2011): **Anonymization of location data does not work: a large-scale measurement study**. Accessed from: http://www.sigmobile.org/mobicom/2011/slides/128-anonymization-slides.pdf

Appendix A: Technical notes on mobile networks and CDR data

Cellular phone systems infrastructure: relevant characteristics for location tracking

The SIM card identifies a specific phone subscriber to the network, whenever the phone is switched on and in range of a cellular mast. The network then tracks the location of the user as they move between cells – and indeed if they 'roam' onto a different network (although roaming is not prevalent among developing country phone users and can probably be ignored for the time being). The network does not however typically create records of movements, only of calls made.

The mobile operator's system network comprises the following main components:

- **Base stations (strictly: Base Transceiving Stations or BTS)**. An individual 'cellular mast' and antenna. By using several frequencies it can handle multiple phones at one time. Several BTSs are grouped through shared control systems (Base Station Controllers or BSC).
- **Mobile Switching Centre (MSC).** Conceptually, this is the network's central hub (although actually a large operator may have more than one MSC). This holds registers of subscribers (and roaming 'visitors' to the network) and logs their calls and other data.

Without very complex triangulation of technical data, it is only feasible to achieve a locational 'fix' from usage data by tagging the call record to the location of the particular cell antenna via which the call was made. In rural areas, cells may cover tens of square kilometres but are much smaller in urban environments. Cell boundaries will not, of course, align to civil administrative boundaries. Therefore locations derived from CDR datasets will be approximations of the caller's actual position, based on the location of the cell antenna that was responsible for handling the call. However, in the humanitarian applications as envisaged, this limitation is unlikely to be problematic.

What is CDR data?

In fact, what is referred to generally as 'CDR data' usually includes two elements: (a) the Call Detail Record (CDR) plus the Call Management Record (CMR) which contains other technical data about the call. However the data of particular interest to us is in the CDR.

CDRs are retrieved and processed by the operator's Business Support System (BSS), so that bills can be generated or prepaid credit accounts debited. The data is also used to analyse network traffic for commercial purposes.

A CDR record is tagged with the phone subscriber's reference number, which will then be joined to the subscriber database to (for example) bill the user. 'Raw' CDR records are therefore not anonymised, although the subscriber's name and other personal details cannot be derived from the CDR alone. The data structure for the CDR is determined by the network technology provider, eg Cisco. CDR data would normally be exported using an XML-based tagging schema specified by the technology provider.

Note that the Haiti analysis, and other studies from the literature review, looked only at the location of the caller, not the location of the call recipient (the 'call destination' or 'endpoint). It is not clear, in any case, how straightforward, technically, it would be to derive the location of the recipient – which may in any case be a phone on a different network, or a landline. It may however be possible to discern local from regional calls.

Appendix B: Processing of CDR data

The following is a brief exploration of technical and organisational approaches to processing of CDR data.

Data processing resources

It is apparent that the CDR-driven methods require, for each instance in a new country and/or context, considerable technical resource for project design, data induction and preparation, processing of analysis (on very large datasets), and interpretation of results. There are probably numerous ways in which these tasks can be structured and they could be distributed between several partner organisations. However, it remains unclear how to define or assess the capacity of potential processing partners or to project manage the entire process chain to achieve predictable outputs. It is however apparent that, as a starting point, expertise in handling and performing statistically valid analysis on datasets with millions of records is a requisite, while understanding of social science principles is possibly also of high importance.

It seems likely that Karolinska Institute has the most relevant experience in processing and analysis of CDR data for humanitarian applications, derived from the Haiti experience. Researchers from Karolinska Institute have set up the Flowminder programme (www.flowminder.org) in Stockholm, with the stated aim of becoming a clearing house for aggregating, analysing and disseminating mobile phone location data to humanitarian actors. It is believed that further funding is being sought to operationalise the proposed services, but actual plans to roll out the service are not known at present.

Aside from Karolinska/Flowminder, workers in several other institutions have apparent experience in CDR-based research: including the University of California Davis, AT&T Labs in Florham Park, NJ and the Department of Electrical Engineering at Princeton University.

Discussions with the GSMA indicate that selection of processing partners for each CDR-derived project, whether pre- or post-disaster, is likely to have a strong interplay with the willingness of the CDR data owners (and of other 'gatekeepers' such as regulators) to sanction and provide access to raw datasets, and on the imposition of processing and sage conditions. Geographical and institutional proximity are both likely to be important factors. It seems likely that a 'local' (that is, same-country) processing partner might often be more palatable to the MNO and other stakeholders than where cross border transfer of data is involved.

MNOs may themselves have the capacity to undertake some data processing and analysis. In any case, they would need to conduct the initial extraction of CDR records from their network admin systems to hand over to a processing partner. This will involve more than trivial effort: CDR data may be archived offline and anecdotal reports (Global Pulse 2012) describe researchers having to delve through boxes of magnetic backup tapes to retrieve data at mobile phone companies' offices. This assumes of course that the companies actually retain the CDR data at all: the New Zealand study suggests that this may not be the case at least for SMS data.

It remains an open question whether MNOs would wish to pseudonymise the data themselves prior to release: probably this depends greatly on the identity of, and agreed procedures with, the processing partner and perhaps on other factors such as humanitarian urgency.

In order to analyse the data and to turn it into actionable information products, the processing partner(s) will need to have access to appropriate software tools and reference datasets. In the Haiti study (Bengtsson et al 2011) software tools used included SQL Server 2005, Microsoft Visual C# 2008, MatLab, and Esri ArcGIS. There was no requirement to use more complex analytics tools for 'sense making' of the data, as the approach to analysis was predetermined. Reference datasets for correlation and outputs visualisation included Digicel-provided network tower and coverage maps, and admin boundary datasets and baseline populations: these latter datasets would normally be part of any humanitarian information Common Operational Datasets collection (as already described earlier in this paper).

Anonymisation of CDR data

Anonymisation (or more likely pseudonymisation, in which the individual's name is replaced by a de-personalised reference) of CDR data can be achieved, to a reasonable degree, by substituting an impersonal code identifier for the user ID before

handing over the dataset for analysis[8]. For most application studies in the literature review, this was typically done by the mobile phone operator company. However, it may still be possible to identify particular users in certain circumstances, by matching data against 'quasi identifiers' (Zang and Bolot 2011). There are mitigation methods that can be applied to reduce if not eliminate this privacy vulnerability. On the other hand it may be judged that the humanitarian benefits of using datasets that have only been pseudonymised at a basic level may justify a small possibility of re-identification of individuals.

In some research studies using CDR data (eg Becker et al 2011) a trusted third party was used to anonymise the data before handing it over for analysis, either as individual records or in aggregated form.

[8] Note that the user ID identifies the SIM card but the CDR record itself does not contain the phone number.

Appendix C: Information product design considerations

Desirable formats for CDR-derived information products

Predictability and ready comprehension of CDR information would be achieved by establishing a standard format for reports and summarised datasets. A recommended format for such reports will be developed through the planed trials process later this year. However, the report contents might include the following:

- Metadata summary: when the data were collected; overall geographical coverage; report version; etc.
- Known sensitivities on data dissemination (eg protection concerns).
- Summary of key findings: up to ten bullet points
- Map depicting key migration patterns as inferred from analysis.
- Constraints and caveats on use of the data: representativeness; coverage gaps; other issues affecting reliability.
- Detailed analysis and findings.
- Optionally: discussion on triangulation with other data.
- Plans for further data acquisition and reporting.
- Contact details for queries.

The structuring of summarised datasets as information products will also be addressed during the trial. It will be important to consider alignment with other humanitarian data standards and approaches, including the *Humanitarian Exchange Language* (HXL) initiative currently being considered by OCHA and other humanitarian agencies.

Mapping and visualisation of CDR-derived data

Because a large proportion of important information in humanitarian crises has a spatial component, visualisation using maps predominates in relief programming and operations. GIS allows data of different types or from different sources to be overlaid to understand readily relationships and patterns. It is likely that location data from CDR records would be displayed and used, in combination with other data, in such thematic maps as decision support tools.

If CDR data were available aggregated to respective mobile phone base stations, it would be necessary for the GIS operator to know the geographical coordinates of the cellular infrastructure, in order to map the data. It is assumed that the data owner would provide these coordinates (eg as latitude/longitude), either embedded in the CDR records themselves or as a look-up table cross referenced to the CDR file. Aggregated data could be provided in any commonplace file format, eg a DBF or Excel file and these could be readily imported into the GIS.

Display of data could be for each cell as a single point (the antenna location) or parsed into approximated cell boundaries. These cell polygons could be generated in the GIS although would be imprecise. It might be more appropriate to associate antenna locations with administrative districts, given that much pre- and post-disaster data is likely to be referenced to the country's admin geography. This would be imprecise because the call may have been made from a village in district X, via a cell mast located in the adjacent district Y; however such distortions may not be crucial.

Call volumes or other variables (eg movements of sub-groups of users) could then be displayed easily though a range of symbolisation options offered by the GIS software.

The Ajala (2006) paper referenced in this document contains a use-case description of the mapping of CDR-derived data in a developing country context.