FlowKit:
Unlocking the power of mobile data for humanitarian and development purposes

FLOWMINDER.ORG

dial Digital Impact Alliance

FlowKit: Unlocking the power of mobile data for humanitarian and development purposes

This paper was authored by Daniel Power, Martin Thom, Jonathan Gray, Maximilian Albert, Sophie Delaporte, Tracey Li, James Harrison, Joshua Greenhalgh, Nick Thorne and Linus Bengtsson from Flowminder.

Acknowledgements

Flowminder would like to thank Kate Wilson, Syed Raza, Laura O'Brien, Melissa Johns, Matt Haikin, Theresa Nyamupachitu and Hitoishi Chakma from the Digital Impact Alliance (DIAL) for their work on this white paper.

FLOWMINDER.ORG

About Flowminder

Flowminder provides insights, tools, and capacity strengthening to governments, international agencies and NGOs. Flowminder is a non-profit organisation composed of practitioners and academics from development, public health and humanitarian sectors. The mission of the organisation is to develop and operationalise new methods and data sources to support decision makers, improve operational efficiency and deliver services in low- and middle-income countries.

Flowminder has pioneered the use of mobile operator data to address development and humanitarian challenges. It uses statistical techniques to integrate traditional data from surveys and censuses, with data from satellites and mobile operators.

Working closely with researchers at University of Southampton UK and elsewhere, Flowminder develops new methods and data to understand the distributions, characteristics and dynamics of human populations, and provides support for decision makers to leverage the operational opportunities of radically improved sociodemographic data.

www.flowminder.org

dial Digital Impact Alliance

About DIAL

DIAL aims to realize a more inclusive digital society in emerging markets, in which all women, men and children benefit from life-enhancing, mobile-based digital services. A partnership among USAID, the Bill & Melinda Gates Foundation, the Swedish government, and the United Nations Foundation, DIAL helps accelerate the collective efforts of government, industry and NGOs to realize this vision.

www.digitalimpactalliance.org

Contents

Executive Summary

While the commercial world routinely uses mobile data to do everything from targeting food purchases to optimizing one's route to work, the humanitarian and development sector lags behind in optimizing service delivery with mobile network data. Despite the fact that 5 billion people are currently connected to the mobile internet, the data generated from cell phone use is still a novelty for most humanitarian and development organizations due to policy concerns about user safety and the lack of turnkey product solutions. Hundreds of humanitarian and development pilots have successfully demonstrated the power of mobile data and its effectiveness in addressing disaster and emergency response, but the routine use of this data still eludes governments, NGOs and multilateral institutions.

To address a key part of this problem, DIAL, GSMA and Flowminder have partnered to create FlowKit, which takes Flowminder's proven, proprietary system and makes it open source. FlowKit promises to turn what is today a specialised practice—using data types produced by mobile network operators (MNOs) to inform public health, urban planning and crisis response—and make it routine. Our vision is to help the sector utilize turnkey products and best practices in a timely and responsible manner for achieving greater efficiency and effectiveness.

FlowKit is a suite of software tools designed to enable access and analysis of mobile data for humanitarian and development use cases. FlowKit supports these goals in three key ways:

1. FlowKit provides humanitarian and development actors with an analytical toolkit developed specifically for their use cases, informed by Flowminder's experience working in this area. FlowKit facilitates collaboration between MNOs and humanitarian and development organizations by breaking down barriers and lowering overhead costs, making it easier for all partners to reap the benefits of forming partnerships. FlowKit provides the technical support and guidance organizations need to install and implement it. Organizations can also install, utilize and customize FlowKit on their own if they have the capacity.

2. FlowKit allows for the secure and compliant access of mobile data behind an MNO's firewall, which helps surmount one of the biggest obstacles to forming partnerships between MNOs and humanitarian and development organizations. In addition, FlowKit includes data quality assurance tools, creating further efficiency gains for analysts.

3. FlowKit is an open source, community-focused software project. All of its documentation, processes and code are openly accessible and available, promoting greater confidence among all stakeholders. As an open source project, FlowKit will continue to evolve as techniques and learnings derived from research are incorporated back into the core tools.

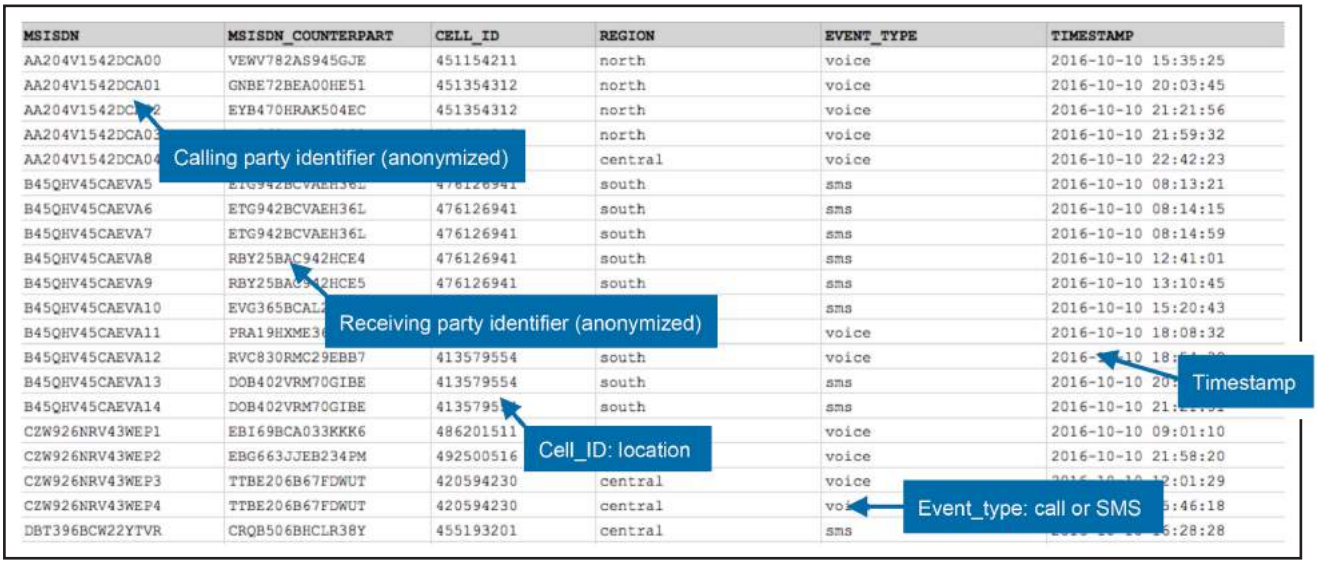
This white paper introduces the key features of FlowKit, providing detailed examples of how these tools assist humanitarian and development actors to improve decision-making and achieve impact. We hope readers will find a use case that is relevant to their own applications. We invite readers to explore the toolkit for themselves or contact us directly at flowkit@flowminder.org for more information.

How mobile data can assist in humanitarian and development efforts

More than 5 billion people are now connected to mobile services worldwide.[1] The data generated by mobile phone use can provide a dynamic, near real-time picture of the mobility and movement of millions of people across entire countries. De-identified mobile phone metadata records can be analyzed to extract insights about where people are, where they are going and the route they took to get there.[2,3] This mobile phone data, in combination with other data sources, can also provide information on people's characteristics, such as the socioeconomic status of individuals and communities.[4] With the right tools, actionable insights can be derived from this data and made available to decision-makers in a timely and accessible way.

For example, understanding the location of displaced people during humanitarian emergencies is a key factor when assessing the scale and impact of an event. This information is also important for the management of relief response, as it informs decisions on suitable locations for on-the-ground assessments for initiating relief assistance.

Call Detail Records (CDRs) are an example of a mobile data type that can be used to provide rapid insights into people's movements, including displacement.[5] CDRs are automatically generated by mobile network operators for billing purposes and created for each pair of SIM cards at the time that a call, SMS or mobile data session is initiated. Importantly, CDRs include an identifier of the cell tower to which the event was routed (usually the closest cell tower), meaning that the movement of SIM cards can be tracked over time and across areas. With appropriate methods to account for the biases in this data, this information can be used to estimate how a population is moving during a period of interest.



Call Detail Records Metadata Example (Source: Flowminder)
Anonymized call detail records (CDRs) provide large volumes of locational data that, with the right tools, can be converted to actionable insights for humanitarian and development decision makers.

[1] The Mobile Economy 2018, GSMA
[2] Unlocking MNO data to enhance public services and humanitarian efforts, DIAL 2018
[3] Incorporating Big Data into Humanitarian Operations 2015
[4] Steele, et al., 2017
[5] Considerations of other mobile data (e.g. from smartphones) are beyond the scope of this paper.

Analyses using CDRs and other mobile data types can be particularly valuable when working in a resource- or time-constrained context, such as commonly found in low- and middle-income countries and in crisis-response scenarios.[6] Mobile data can also be included in more sophisticated analyses, such as those incorporating household survey data and/or geospatial data.[7] If sufficient time and resources are available, advanced statistical modelling techniques can be used to produce a more detailed and accurate picture of a situation.

Considerations for analysts using mobile data

As with all data types, it is important to consider the limitations and biases that exist when performing analyses.

Firstly, mobile data is biased if used to represent the entire population of a country (selection bias). Anonymous data appearing in an operator's data is, for obvious reasons, coming from people who use a mobile phone and who make calls that involve a SIM card. This typically means that the youngest and oldest segments of a population, as well as those in the lowest socioeconomic strata, generate much less data than they would if mobile phone use was spread equally throughout society. Very often, women and other vulnerable groups are also under-represented.[8] If these under-represented groups behave differently, and such differences are not adjusted for, bias will occur.

Secondly, the assumption that one SIM card corresponds to one person may be less appropriate in communities where it is common for groups of family members or friends to share a phone, and for individuals to own multiple phones, often with SIM cards from different operators. This can be particularly relevant to analyses where it is necessary to know the sequence of movements of individuals, such as in epidemiological studies.

Thirdly, the spatial and temporal resolutions of the data are limited by frequency of phone usage, tower coverage, and the density of cell towers. All of the above limitations should be considered during the analysis of the data and interpretation of the results. If done so appropriately, then mobile phone data can be an invaluable source of information during a humanitarian crisis and for development applications.

[6] Bengtsson, et al., 2011
[7] Wilson, et al., 2016
[8] GSMA Case Study on Bridging the Gender Gap in Mobile Access

Case study

Using CDR data to support relief interventions, Hurricane Matthew 2016

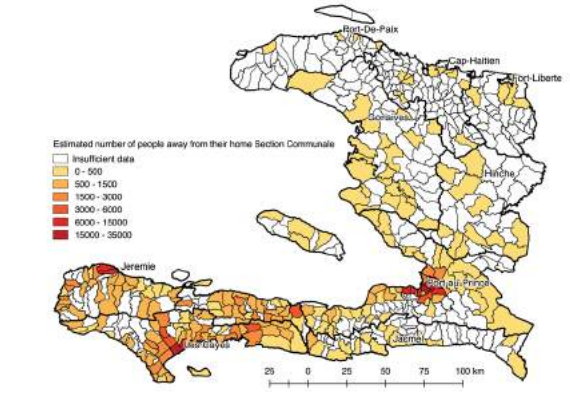
On October 4 2016, Matthew, a category 4 hurricane, made landfall over the southwest of Haiti, causing massive damage to the Grande Anse and Sud regions. Hurricane Matthew took the lives of more than 500 people and caused more than EUR €1.8 billion in damage.[9] Within a week of the hurricane, OCHA estimated that more than 175,000 people had been evacuated to temporary shelters and many of the more than 1.4 million affected by the hurricane were displaced to unknown locations.

More than 400 humanitarian operations took place in the aftermath of the hurricane.[10] The World Food Programme (WFP) was one of the organizations that participated in the response. Flowminder helped WFP collaborate with the Haitian Mobile Network Operator Digicel with the aim of incorporating insights from mobile data to inform relief efforts.

Before Hurricane Matthew struck, Haiti was already dealing with an ongoing cholera epidemic. The situation was exacerbated by the hurricane, as water supply infrastructure, sanitation and several existing cholera treatment centers sustained heavy damage.[11] The presence of cholera increased the need to identify displacement, with a particular emphasis on understanding where people had come from, as this information was used to inform relief management efforts on containing the disease among a highly vulnerable population. Flowminder was able to support WFP in distilling this information from Digicel's mobile data, information that would have been difficult and costly to obtain through other means, particularly in the post-disaster context.[12]

24 October 2016, location of people away from their home Section Communale

(out of those living pre-hurricane in Grande Anse, Sud and Nippes only)



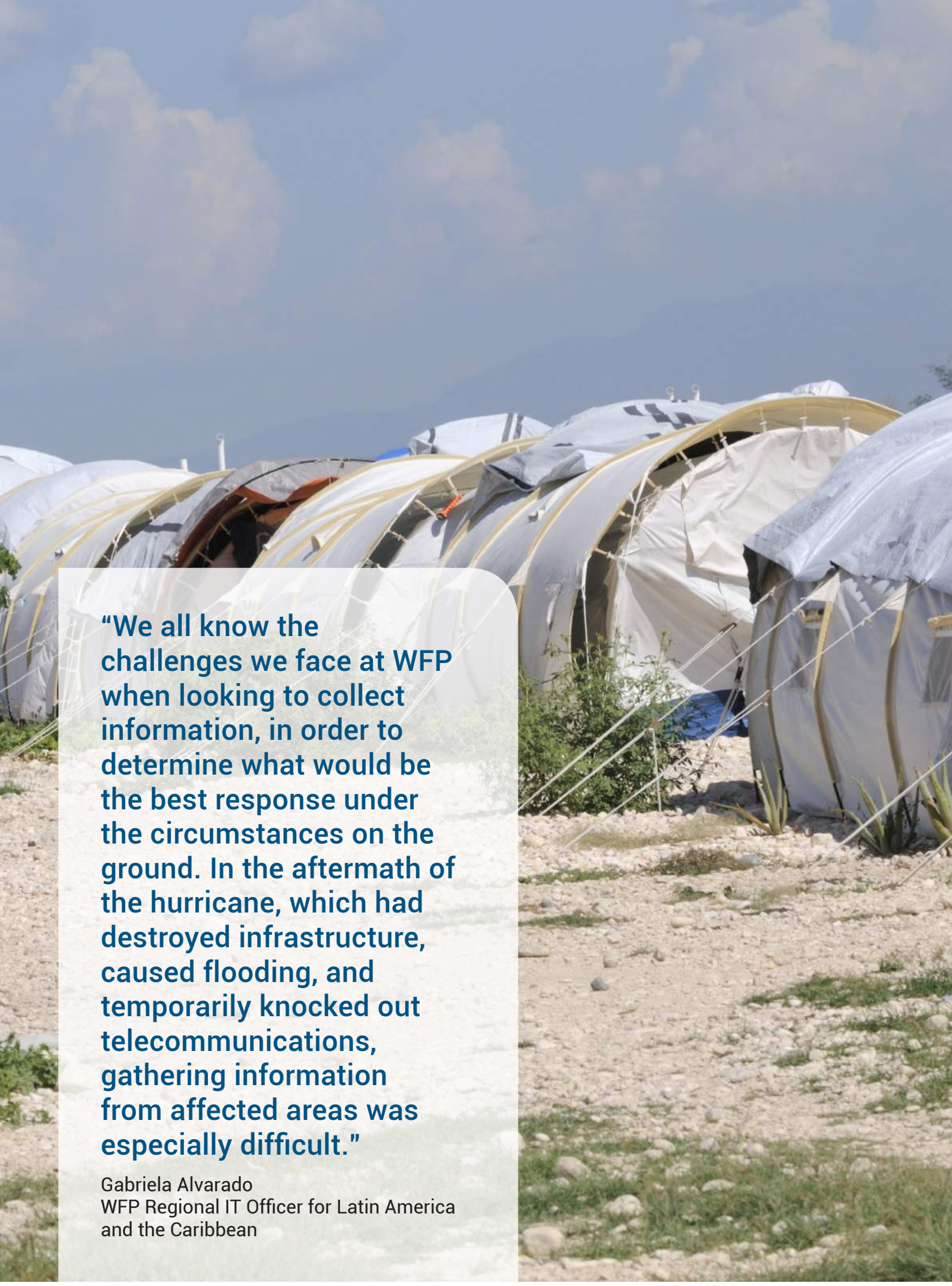
FLOWMINDER.ORG Digicel

Section Communales are left blank where insufficient data is available.

Hurricane Matthew: Estimated population movement, October 24, 2016. Showing movement of people away from Section Communale.

Source: Flowminder, taken from the Humanitarian Data Exchange

[9] ReliefWeb
[10] Hurricane Matthew OCHA Report
[11] UNICEF report
[12] World Food Program's MVAM Blog



"We all know the challenges we face at WFP when looking to collect information, in order to determine what would be the best response under the circumstances on the ground. In the aftermath of the hurricane, which had destroyed infrastructure, caused flooding, and temporarily knocked out telecommunications, gathering information from affected areas was especially difficult."

Gabriela Alvarado
WFP Regional IT Officer for Latin America and the Caribbean

FlowKit unlocks mobile data for humanitarian and development decision-making

FlowKit is a suite of software tools that help humanitarian and development organizations access and analyze mobile data, allowing them to use this data to develop insights and improve decision-making. FlowKit is targeted at the distinct needs of humanitarian and development practitioners in three key ways:

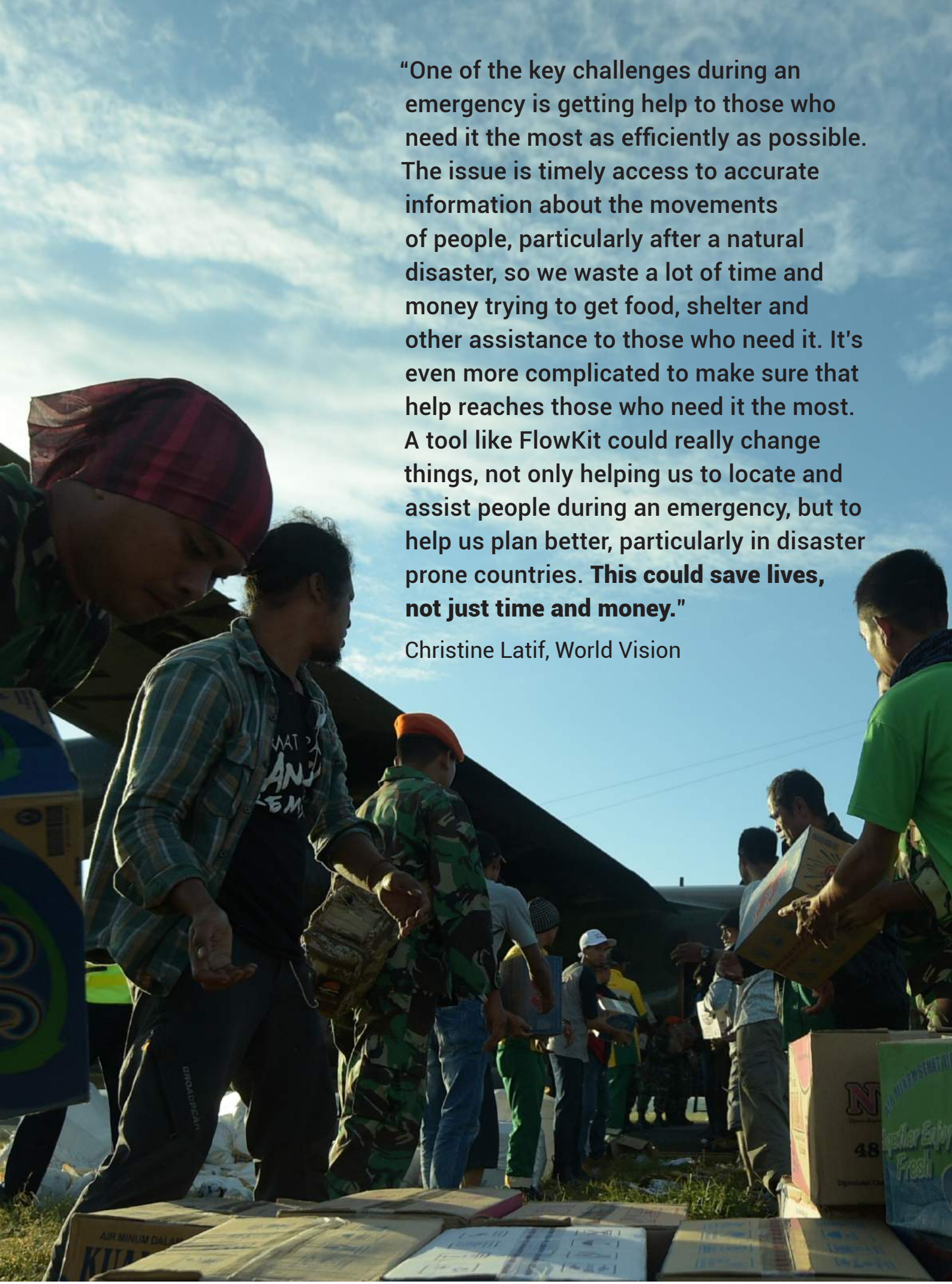
1 **A supported toolkit for mobile data analysis:** FlowKit provides a mobile data analysis toolkit informed by the Flowminder team's 10+ years worth of experience producing insights from mobile phone data to support the humanitarian and development communities. Because FlowKit has been developed with a strong understanding of the needs and priorities of collaborating partners—humanitarian and development organization and MNOs—it is able to facilitate collaboration and help form strong partnerships. It provides the necessary features and support to allow all parties to engage in an easy, low-cost manner. Installation and setup of FlowKit can be done independently. Alternatively, Flowminder offers additional support and guidance to organizations that do not have the skills and capacity to use FlowKit on their own. The implementation of this toolkit can be modified and customized according to the needs and constraints of a range of different end users.

2 **Secure and compliant data access:** FlowKit facilitates secure and compliant data access. Given that complications around security and privacy tend to be the single biggest blocker to this type of data being used by humanitarian and development organizations, this is a key value-add in facilitating these engagements. FlowKit also provides processing with built-in quality assurance, lowering overhead costs for users and smoothing the path to data analysis.

3 **Open source:** FlowKit is released under an open source license[13], thus promoting the transparency of its techniques as well as future sustainability by encouraging additional development and contributions from others working on similar challenges. The software will continue to evolve as research and learnings from Flowminder and those participating in the new FlowKit community are incorporated back into the core tools.

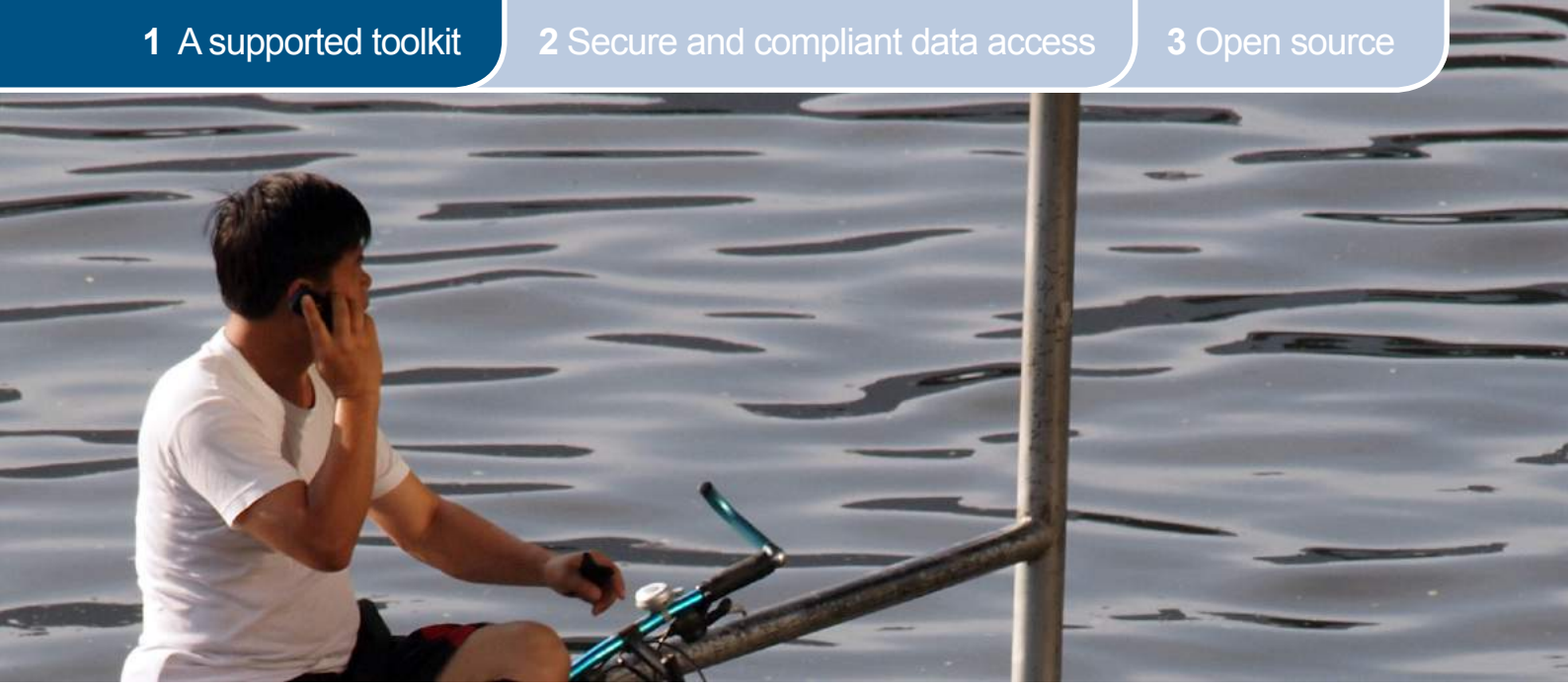
---

[13] FlowKit was released under a Mozilla Public License, version 2.0 (MPLV2). More details about the license can be found here: https://www.mozilla.org/en-US/MPL/2.0/



"One of the key challenges during an emergency is getting help to those who need it the most as efficiently as possible. The issue is timely access to accurate information about the movements of people, particularly after a natural disaster, so we waste a lot of time and money trying to get food, shelter and other assistance to those who need it. It's even more complicated to make sure that help reaches those who need it the most. A tool like FlowKit could really change things, not only helping us to locate and assist people during an emergency, but to help us plan better, particularly in disaster prone countries. **This could save lives, not just time and money.**"

Christine Latif, World Vision



1 A supported toolkit for mobile data analysis

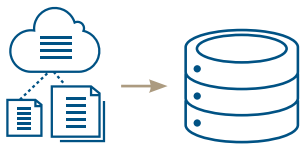
A flexible toolkit to meet diverse users' needs

FlowKit's powerful library of features and tools have been developed specifically for the use cases of humanitarian and development actors. Access to the library is facilitated through FlowKit's application programming interface (API),[14] which provides an auditable and secure method for multiple users to access analytical outputs generated by the tools provided by FlowKit.

Installation of FlowKit behind an MNO's firewall can be done by MNO engineers or a third party that has met the MNO's security requirements and has the necessary permissions. Flowminder can provide advice and/or directly support installation and setup. For users wanting to explore or test

What is FlowKit?

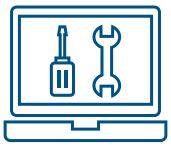
FlowKit automates the retrieval and validation of new sets of results for defined time periods.



Mobile data Database

Ingestion and QA

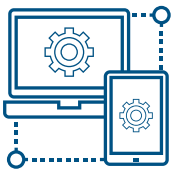
With its additional tools, analysts can gain access to mobility insights, characteristics and actionable results.



Tools

Analysis and scheduling

For most applications, access to the results is determined by controlled and audited interactions through the API.



API

Authentication and controlled access

[14] An API is a set of procedures used by computer programs to request information from another system. It serves as an interface and facilitates their interaction. It provides developers with standard commands to perform operations without having to deal with the complex internal settings of the toolkit.

FlowKit's capabilities outside of this environment, FlowKit comes complete with demonstrations and test data that can be run on standard computers, without the need for real mobile data.

Once installed, FlowKit provides a versatile suite of tools that can be tailored to meet the needs and capabilities of its users. There are three typical user groups of FlowKit:

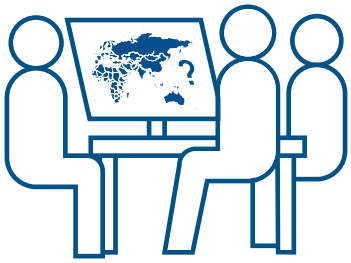
The primary users of FlowKit are **humanitarian and development actors whose needs are met by FlowKit's automated outputs**. Many of FlowKit's built-in tools have been designed to place low technical and analytical demands on users. With appropriate support, these users do not have to interact with the software at all. A specialized technical team from the humanitarian or development organization, the MNO, or an organization such as Flowminder can install and configure FlowKit according to the user's specifications. Once validated, information can then be disseminated and incorporated into decision-making.

The second user group is **data analysts, usually in humanitarian and development sectors**, who interface with FlowKit's API to carry out their analyses. This API provides access to relevant analytical tools while giving them the flexibility to refine their investigations. FlowKit supports these analysts' need to interact with data as they develop a better understanding of a particular context and incorporate relevant information from other data sources.

The third class of users supported by FlowKit is **data scientists with high technical expertise**, who require deeper understanding of specific questions beyond the tried-and-tested algorithms and outputs provided by FlowKit's API. They can take advantage of the existing data processing and quality assurance framework and supplement it with their own bespoke tools. FlowKit provides efficiencies to these users by reducing the overhead in setting up a data pipeline and establishing a data quality control system.

SCENARIO A

Crisis Preparedness Programme Manager at an NGO



In time of crisis, humanitarian and relief agencies' biggest challenge is to determine where people are and to where they are moving.

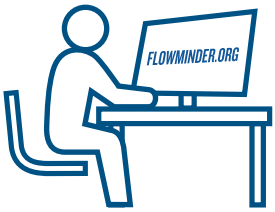
Programme Managers of those agencies require accurate data as fast as possible to mobilise their relief teams to priority areas and support those in need.

As part of their crisis preparedness plan, NGOs can collaborate with MNOs and Flowminder to install FlowKit and let it start analysing data.

Additionally, in the time between installation and crisis, FlowKit is still available to perform other analyses (see scenario B)

No direct agency interaction with FlowKit is necessary. Programme Managers can call upon Data Analysts to identify the outputs they would require at the time of a crisis. These will then be made ready when the crisis hits so that the field team can focus on initial response logistics.

A lot of the work is done at the preparedness phase. Data Analysts prepare the relevant analytical functionalities, based on parameters defined with the NGO.

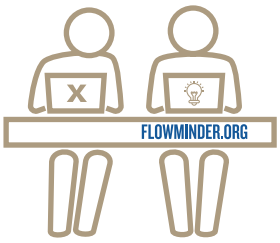


In the immediate aftermath of a crisis, and assuming the appropriate preparation has been done, there is only a small amount of work for analysts to do to assess the accuracy of the post-crisis information and then disseminate the results in a clear way.

Insights are sent near real-time to the Programme Managers in the format they need, to a website or other appropriate channel used by the relief teams, enabling rapid and effective interventions to areas affected.

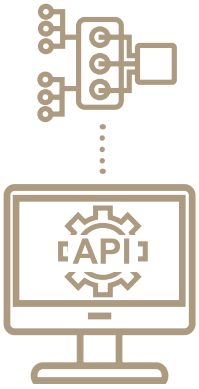
SCENARIO B

Data Analyst in a humanitarian and development agency



Data analysts who can see the benefit of the mobility and characteristics information derived from mobile phone data can interact with FlowKit to extract aggregated data on the locations and movements of populations, and other relevant metrics. This data can be useful in itself or in combination with other data available to the analysts.

MNOs who want to explore their own data can also interact with FlowKit to create data that has broader value for their business or community.



Once FlowKit has been installed behind the MNO firewalls, analysts can interact with the data through an API, thus allowing access whilst still preserving the appropriate security and privacy controls.

The API is an interface that gives restricted, audited and controlled access to the algorithms and methods that are built into FlowKit. The analysts can extract aggregated data such as information about locations, movements of the population and subscriber characteristics and behaviour.

These users can interact with the system to carry out the analyses they need, when they need them. The data extracted can be analysed on its own, used for inputs in further calculations, or combined with other data sources to generate custom analyses.

FlowKit is built to continually ingest new data. Since mobile data is of greater temporal resolution than traditional survey data, the analysts are able to return repeatedly to this new information and seek out new conclusions. Analysts can respond alongside and in time with the evolution of the situation they are evaluating.

The FlowKit algorithms and methods are not fixed in stone. Analysts can change some of their parameters in their data queries (e.g. timeframes, locations…) using the FlowKit algorithms to extract the most relevant reports and update their analyses.

SCENARIO C

Data Scientists requiring bespoke analyses

On some occasions, it may happen that a data scientist interacting with the FlowKit API realises they need an additional set of functionalities over the ones built into FlowKit to carry out their analyses.

Once the data scientist has had the appropriate security and privacy training, to the satisfaction of the authentication controller, they can be granted the privilege of more direct access to the data and library tools within FlowKit.

By gaining direct access to the library of tools and data within FlowKit, those data scientists can build upon existing code, algorithms, quality checks, and kick-start the development of the enhancements that they wish to introduce; even to the stage of developing and incorporating their own additional tools.

In the future, these new functionalities could be made available to the general users (see scenario B) through the API. This user type can contribute to the ongoing development of FlowKit.

Analytical features to meet the needs of humanitarian and development actors

FlowKit provides access to insights about subscribers' characteristics, network usage, mobility patterns and more. These metrics, which are useful in and of themselves, are particularly powerful when used as the building blocks of analyses and enriched further with contextual and supplementary data. The table below gives an overview of FlowKit's core analytical features, broken down by insight category, and some potential applications of these tools in humanitarian and development interventions.

Insight category	Example FlowKit metrics	Example application
Location	Meaningful locations (e.g., home, work) for users over specified time periods, most frequent location, last location in a time period	Dynamic population mapping, post disaster displacement monitoring
Mobility	Origin destination matrices (see spotlight on p. 17), or flows between two time periods; multiday trajectories; flows between meaningful locations; population weighted opportunities model	Identifying unusual patterns of mobility post disaster, commuter flows
Subscriber behavior	Radius of gyration, nocturnal calls, active periods, call duration statistics	Poverty mapping, subscriber profiling, disaster early warning
Network statistics	Voronoi and viewshed coverage maps, network activity by location, number of new subscribers in a specified time period	Population mapping, network outage monitoring, post disaster network recovery monitoring
Subscriber characteristics	TAC codes, handset models, subscriber community detection	Subscriber profiling, poverty mapping, informing survey methods

This is a high-level overview of the features provided by FlowKit's library and how they relate to user requirements. Some of these features are expanded in more detail in the table in the **appendix on p. 27** and at **FlowKit's documentation website**, where it is regularly updated as new insights are added or made available under FlowKit's API.

Feature in focus

Origin destination matrices

Origin-destination (OD) matrices, one of the key features available in the FlowKit library, are a commonly used means for representing population movement between two time points. For specified locations (e.g. administrative regions in a country), OD matrices show the number of people who moved between locations between time points. The key considerations in calculating an OD matrix from mobile phone data are: (1) the spatial resolution and meaningful locations (e.g. home locations[15]); and (2) the temporal resolution, meaningful time points for the question of study (e.g. before and after an event). Once these parameters are chosen, one can calculate the locations of each subscriber during the times of interest and aggregate this information to derive the OD matrix.

The choice of spatial and temporal resolution depends very much upon the application. The method or technique used to calculate people's locations similarly depends on the specifics of the questions being answered. FlowKit enables users to easily specify appropriate parameters and select appropriate methods for calculating locations. The table below illustrates how OD matrices can be used in different applications:

Application	Type of movements of interest	Temporal resolution	Locations of interest	Use cases
Disaster response	Unusual movements that may indicate displacement	Days or weeks	Long-term/short-term resident location before/ after disaster	Displacements after Gorkha earthquake (Nepal, 2015) — see case study p. 19 Hurricane Matthew (Haiti, 2016) — see case study p. 7
Infectious disease control	Regular and irregular movements	Days to years	Areas receiving population groups from other areas with ongoing transmission	Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data Containing the Ebola Outbreak — The Potential and Challenge of Mobile Data
Commuting patterns	Regular movements	Daily patterns	Home and work locations	A trip to work: Estimation of origin and destination of commuting patterns in the main

Example applications of OD matrices in differing applications

[15] A necessary simplifying assumption when constructing OD matrices is to assign each user to a single location over a specified time period. This is often referred to as a "'home location", but can be representative of any significant location for that user (e.g. place of work).

Meeting the needs of mobile network operators

The development of FlowKit has been strongly informed by Flowminder's experience in forming strong relationships with MNOs and collaborating to achieve social impact. MNOs wishing to engage with humanitarian and development organizations must provide internal resources across multiple areas of business. Lack of internal MNO resources can sometimes be a blocker in both establishing partnerships and meeting objectives once collaboration has been established. This applies especially for MNO resources for data management, and particularly in post-crisis situations where an MNO may need to manage impacts to its own infrastructure, as well as supporting humanitarian initiatives. FlowKit enables MNOs to achieve the benefits of partnership by providing tools that facilitate the management, processing and analysis of mobile data. This reduces the overhead for MNOs in supporting humanitarian and development agencies.

Installation of FlowKit also gives MNOs access to FlowKit's tools. Many of the analyses that are developed for humanitarian and development applications are also relevant for MNO use cases. FlowKit is built using a modular architecture that allows MNOs to develop custom code to communicate with FlowKit based on their specific needs. Similarly, FlowKit can incorporate graphical interfaces to enhance analysis, with user-specified choice of analytical ecosystem, e.g., tools developed in the programming language R. FlowKit comes with clear and comprehensive documentation, allowing ease of use and enhancing the ability of the MNO to monitor analysis activities.

Case study

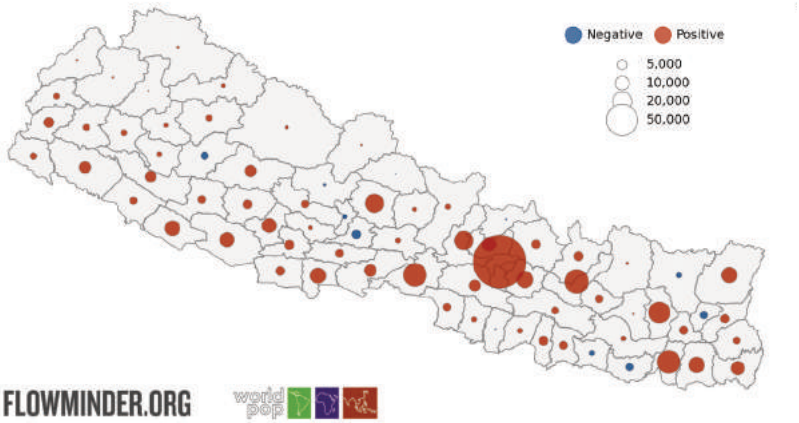
Building from past experience – 2015 Gorkha earthquake, Nepal

FlowKit provides a mobile data analysis toolkit informed by the Flowminder team's 10+ years worth of experience producing insights from mobile phone data to support the humanitarian and development communities. For example, in the field of disaster response, Flowminder's analyses supported operations in the 2010 Haiti Earthquake, the 2013 typhoon in Bangladesh, the 2015 Nepal Earthquake, the 2016 Haiti Hurricane Matthew (see case study on p. 7) and the Ebola response in West Africa.

The 2015 Gorkha earthquake in Nepal caused massive damage and displaced an estimated 390,000 people.[16] Flowminder worked closely with the Office for the Coordination of Humanitarian Affairs (OCHA) and Nepalese mobile network operator Ncell to rapidly provide humanitarian actors with actionable insights on population displacement, which could not have been produced by other means at similar timescales. In the first few days following the earthquake, agency requests to Flowminder for information were primarily focused on understanding the size of population movements in the most affected areas (such as the Kathmandu Valley) and where people were moving to. In the weeks and months following the earthquake, agencies requested information indicative of recovery, particularly the rate at which displaced people were returning to affected areas.

Flowminder and mobile network operators, TeliaSonera and Ncell were awarded the Global Mobile Award for Mobile in Emergency or Humanitarian Situations in February 2016, recognizing the value that mobile phone analyses bring to crisis preparedness planning.

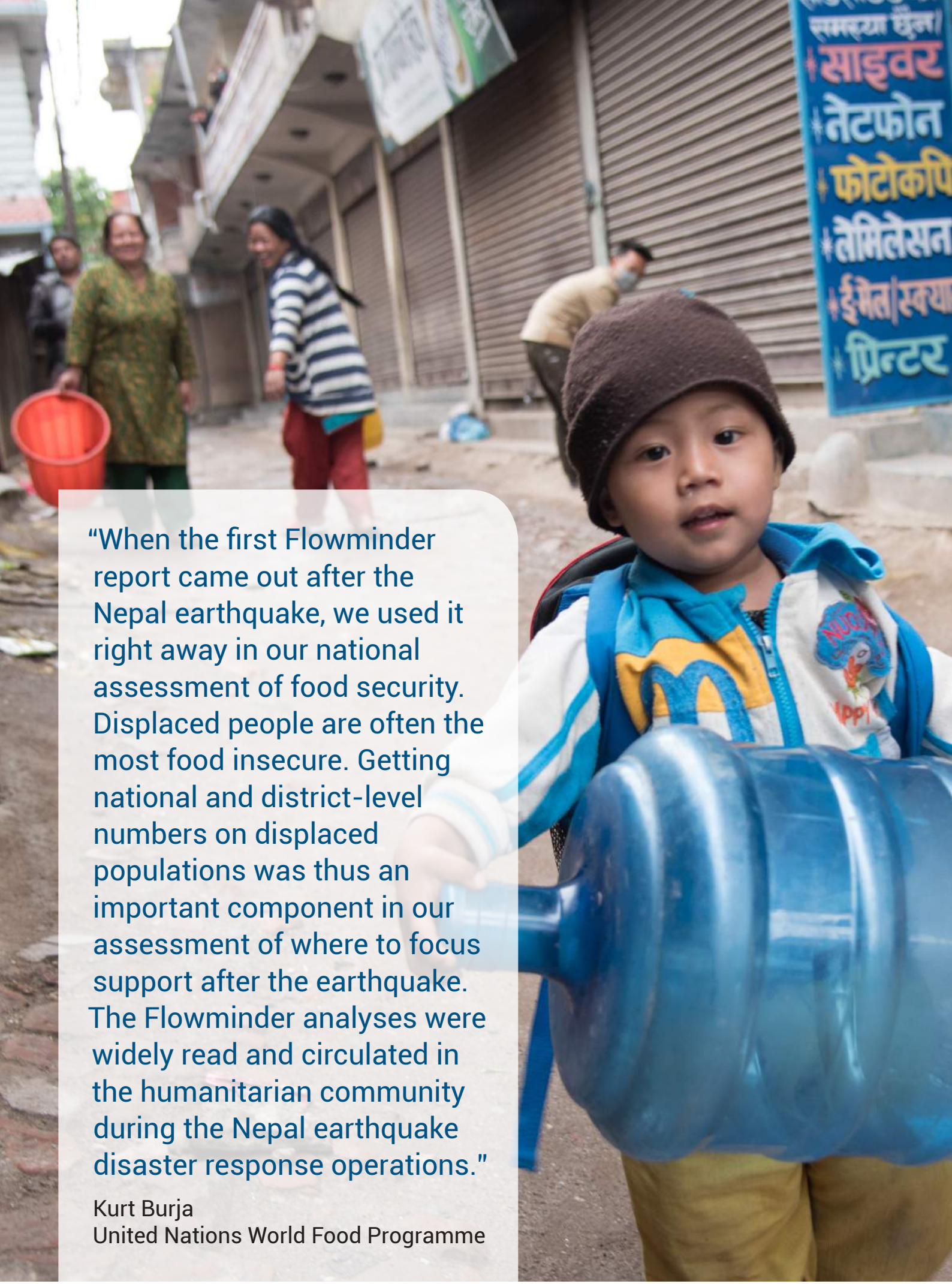
Above normal inflow to each district
(negative numbers indicate less incoming people than expected)



Source: Flowminder, taken from WorldPop

Watch the story here: https://www.youtube.com/watch?v=hY9g0kPi8BA

[16] Wilson et al. 2016



"When the first Flowminder report came out after the Nepal earthquake, we used it right away in our national assessment of food security. Displaced people are often the most food insecure. Getting national and district-level numbers on displaced populations was thus an important component in our assessment of where to focus support after the earthquake. The Flowminder analyses were widely read and circulated in the humanitarian community during the Nepal earthquake disaster response operations."

Kurt Burja
United Nations World Food Programme

2 Secure and compliant data access

The biggest challenge for MNOs and humanitarian and development actors interested in engaging with each other is in overcoming the hurdles in providing the non-MNO party with secure, compliant access to sensitive data. Access depends upon MNO confidence that data sharing can take place in a manner that meets relevant data protection legislation and does not expose the MNO to unnecessary risk.

A first step for any humanitarian and development organization working with an MNO is ensuring that sufficient security controls, systems, processes and documentation are in place. Organizations must demonstrate that data processing will be fully compliant with current legislation, as well as any internal requirements by the operator. Demonstrating success in this area typically includes meeting the GSMA best practice guidelines specifying that all processing will be physically carried out on the operator's own premises.[17]

FlowKit runs *behind* an MNO's firewall, facilitating compliance with data privacy regulations and GSMA guidelines.[18] This ensures that the MNO can control data access, preserve commercial confidentiality and protect its subscribers' personal data. By providing a trusted, secure method for data access, FlowKit simplifies the challenge of facilitating compliant access to this sensitive data.

FlowKit is set up so that de-identified mobile data lives and processing takes place behind an API, through which analysts can access spatially aggregated outputs over an encrypted connection. As an additional security measure, all user activity through the API is logged, providing a record of who has accessed which data, as well as any unauthorized access attempts.

Mobile data that contains information on an individual's location history is classified as *personal* data, even after all personal identifiers have been anonymized.

Data protection legislation exists at the national and global-regional levels (e.g. EU GDPR[19]) to protect the privacy of individuals. Failure to comply with these laws can result in costly penalties, so operators that provide data access to an outside party need to be confident that access is compliant.

Data preparation and quality assurance

Humanitarian and development organizations may have limited resources and specialized skills to manage large-scale data processing activities. Necessary competencies include the ability to efficiently work with a large volume of data and skills in data quality assurance, as appropriate for the analyses being conducted. In addition to promoting secure access to data, FlowKit provides tools that support the accessing party in processing and organizing data prior to analysis. FlowKit also provides quality assurance measures. Given the volumes of data involved, these can be considerable efficiency gains and can significantly reduce overhead costs for both the MNO and the humanitarian or development organization.

[17, 18] GSMA guidelines on the Protection of Privacy in the Use of Mobile Phone Data
[19] European Union General Data Protection Regulation

Case study

Partnership between Flowminder, Vodafone Ghana and Ghana Statistical Services

A data-sharing agreement was established for an ongoing project in which Flowminder is working with Ghana Statistical Services (GSS) to incorporate insights from de-identified mobile phone data, which is provided by Vodafone Ghana, into the production of Ghanaian national statistics. This work is being funded by the Vodafone Foundation and Hewlett Foundation. In order for such an agreement to be viable, it was necessary for GSS and Flowminder to obtain approval from the Data Protection Commission, the national regulatory body in Ghana to ensure compliance with Ghanaian data protection legislation. Proof of compliance with U.K. legislation and the European Union's GDPR was also required. Absolute clarity about the roles and responsibilities of each party with respect to data governance, as well as agreement about the technical aspects of data security, were also key elements that led to a contract being successfully finalized.

Feature in focus

User authentication and access control

User authentication and access control to preserve subscriber anonymity is facilitated through FlowKit's API, which uses access tokens (JSON Web Tokens[20]) to provide fine-grained control for each user for both the types of outputs and the spatial aggregation level a user can access. These access tokens expire after a period of time, limiting exposure in the event of a credentials leak. Access tokens are created by an authentication management system, which runs independently of the main FlowKit server. This enables the isolation of authentication from data handling, supports conducting secure analyses in air-gapped deployments and facilitates centralized management of access by administrators.

[20] https://jwt.io

Feature in focus

Data preparation and quality assurance

Data preparation

FlowKit utilizes a PostgreSQL Database backend, a well-maintained, mature and highly efficient open source database system with strong support for enriching mobile data with geographic information system (GIS) sources provided by the PostGIS extension. FlowKit introduces a specifically designed database schema focused on the analysis of mobile data types. It makes use of a range of PostgreSQL features to increase the efficiency of data queries, including partitioning data by day, index creation, and the clustering of tables, so as to make disk access as fast as possible.

FlowKit provides the ability to specify various aggregate metrics that will be computed and stored as soon as the correct data becomes available.

FlowKit provides the ability to specify various aggregate metrics that will be computed and stored as soon as the correct data becomes available. This means that analysts do not need to configure and initiate computation and can more rapidly access pre-computed metrics. This feature is particularly useful in the context of disaster response, as relevant aggregates can be pre-populated in the run up to the onset of the event.

As a further efficiency, FlowKit provides a caching infrastructure. This ensures that the results of queries of the mobile data that are used often and are expensive to compute will be stored to disk for instant access. The cache infrastructure is also useful in multi-analyst environments, as it reduces the need to run repeated queries by different users.

Quality assurance

Work is ongoing to enhance FlowKit with an automated data loading, or ETL (extract, transform and load) process to convert de-identified mobile data into analysis-ready data within the FlowKit tool. As part of this process of ingestion, a suite of quality assurance (QA) checks can take place. FlowKit allows for the easy specification of QA checks, wherein example checks are provided and individual users can create their own checks to address their specific needs or interests. QA checks fall into two categories:

1. **QA checks that take place before data is ingested into the database.** These checks ensure that new data satisfies certain pre-defined specifications. This enables analysts to rapidly detect any failure to meet these requirements so that mitigating actions can be taken as quickly as possible. Data specifications are negotiated by the FlowKit user and the MNO prior to data transfer. Certain “core” aspects of this specification are always required. These include a subscriber identifier, a timestamp and a location identifier.
2. **QA checks that take place after data has been transferred to the database.** Once data is loaded, FlowKit provides technicians with descriptive statistical analyses and related tools that increase the speed and probability of detecting problems. For example, these second stage QA checks can measure call volumes, enabling comparison of volumes between days (e.g., large differences may indicate missing data); count the number of “never before seen” subscribers, with notifications if this exceeds a user-defined tolerance; and detect variation in cell tower locations, indicating errors in tower labelling.

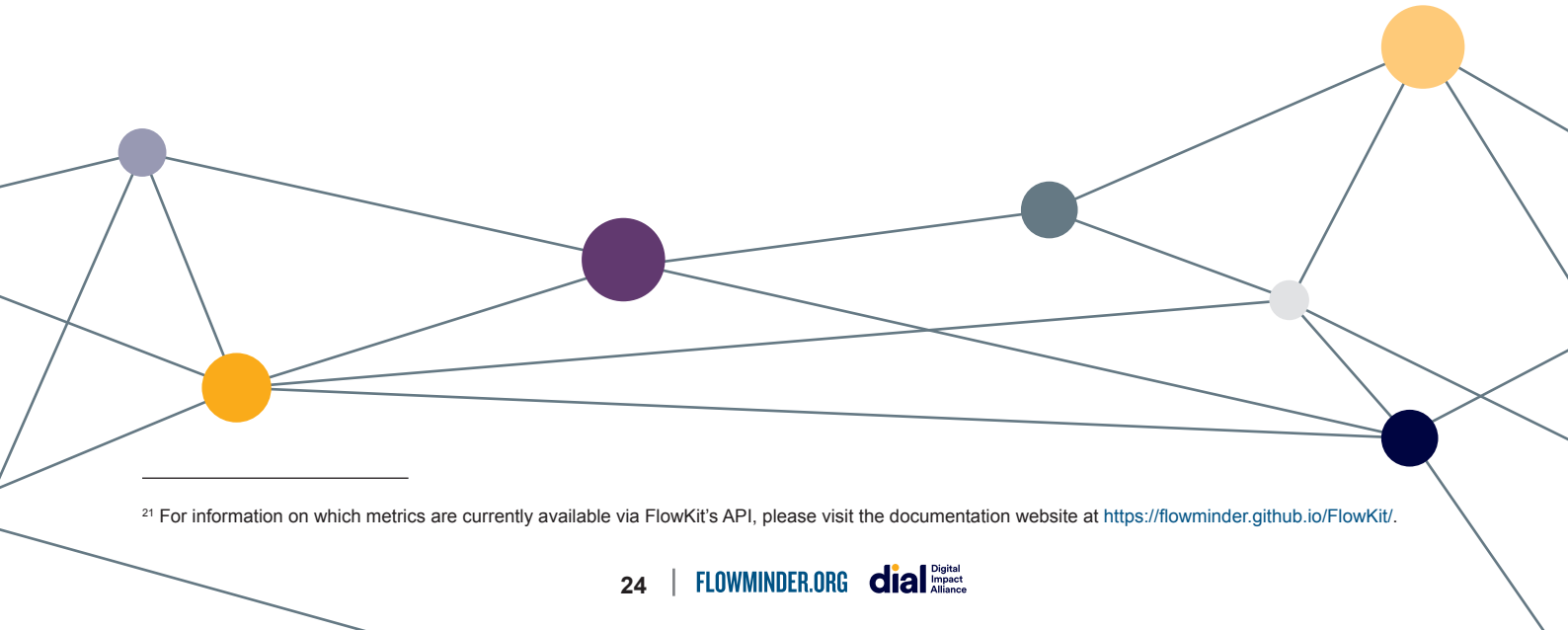
3 Open source

As an open source software project, FlowKit is community focused, and all documentation, processes, and code are accessible and available both without charge and with permission for others to modify and improve. Because the source code is publicly available, every manipulation carried out on the data can be understood and scrutinized by interested parties. FlowKit's supporting documentation on functionality can be independently verified by examining the workings of the code. This transparency promotes confidence among the public, MNOs, governments, and the humanitarian and development communities across all activities, including data handling and protection of subscriber privacy, compliance with relevant legislation, and suitability and appropriateness of processing methods and analytical approaches.

Additionally, as FlowKit is built using a community model, with supporting documentation and demos, it is easy for any interested party to download the tool for inspection and testing. This enables potential users to rapidly assess the suitability of and build confidence in the software for their own needs before committing further resources or deploying on real data. This lowers barriers to entry for humanitarian and development actors, as well as other potentially interested groups, including businesses, software developers, academics and students.

The exposed nature of the implementation of a particular method contributes to collaboration, extensibility, transferability, reproducibility and even standardization of algorithms across platforms in proportion to the amount of engagement received. As the software is increasingly deployed, a rapidly growing community will likely share their implementation experiences and work together to build an open body of knowledge on how the software can be used to address questions and solve problems.

FlowKit is still under active development, both by Flowminder and through open source collaboration, and the number of methods and tools being made available through the API increases on a regular basis.[21]



[21] For information on which metrics are currently available via FlowKit's API, please visit the documentation website at https://flowminder.github.io/FlowKit/.

"FlowKit is the first open source toolkit that I know of to offer a standard way to deal with CDR or mobility data, so the simple fact that this tool is available and it's open source is already a huge deal for me."

Dunstan Matekenya
Data Scientist, DECVP
Development Data Group at the World Bank

Ongoing activities

FlowKit development is directly supported by DIAL and GSMA. Support from DIAL has enabled Flowminder to develop FlowKit's API and authentication manager, improve the codebase, enhance performance and release FlowKit as an open source project. FlowKit will continue to grow and evolve over the coming months with the addition of new insights, additions and enhancements to the API, stronger connectivity to other data platforms, and support for more data analysis environments. Thanks to support from GSMA, 2019 will also see the release of a modular CDR QA suite, coupled with powerful tools for accelerating and automating critical post-disaster insights. FlowKit continuously develops as a consequence of Flowminder's ongoing project work. As a matter of course, all developments take place with privacy in mind as part of the design.

For more details on upcoming FlowKit developments and an up-to-date view of where the toolkit is headed, we encourage you to visit the roadmap [http://flowminder.github.io/FlowKit/developer/roadmap].

How to collaborate towards FlowKit's development

We welcome contributions and discussion on how to develop FlowKit, feedback on what works well, and suggestions for areas for improvement. To join our open source community please visit: https://github.com/Flowminder/FlowKit/blob/master/README.md.

How to get started using FlowKit

You can explore our online materials on GitHub, contact us for more information or arrange a demonstration by emailing us at info@flowminder.org.

Appendix

This table provides more detailed information about some of the FlowKit features. For a more comprehensive list, please visit the FlowKit documentation website at https://flowminder.github.io/FlowKit/.

Metric	Description	Insight categories
Modal location	Mode over most frequent location in a day, or last location in a day for multiple days. Often used as a 'home' location.	Location
Origin destination matrix ('flows')	Count of subscribers with a different modal location in two time periods.	Location, Mobility
Most frequent location in a day	Location a subscriber had the most interactions in a day.	Location
Last location in a day	Location where a subscriber had their last interaction in a day.	Location
Isaacman 'meaningful location' clustering, labelling	Assigns a subscriber a number of significant locations (e.g. work and home), based on their usage patterns for clusters of towers which are near one another.	Location
Location event counts	Count of total interactions in a time period broken down by location	Location, Network
Location introversion	Proportion of interactions in an area which were between subscribers in the same area	Poverty mapping
Location extroversion	Proportion of interactions in an area which were between subscribers in different areas	Poverty mapping
Unique subscribers per location	The distinct set of subscribers who were active in a location during a time period	Location
Total network objects	The number of cell sites or cells observed in use during a time period, broken down by area	Network
Aggregate network objects	The number of cell sites or cells observed in use during a time period, broken down by area and aggregated by time.	Network
Raster stats	Aggregation of GIS rasters to areas	GIS, Location, Mobility, Population mapping, Poverty mapping
Network infrastructure distance matrix	Crows flight distance between all cellular sites or cells	Network
Gridded infrastructure	Mapping of cellular sites or cells to a standardised gridding of a country	Network, Location, Mobility
Viewshed coverage mapping	Coverage maps for cellular infrastructure generated using viewshed modelling	Network
Voronoi coverage mapping	Coverage maps for cellular infrastructure generated using Voronoi tessellation	Network
Radius Voronoi coverage mapping	Coverage maps for cellular infrastructure generated using Voronoi tessellation constrained to a radius	Network

Metric	Description	Insight categories
K-means location clustering	K-means clustering of sets of points, for example locations of network infrastructure	Network
Dbscan location clustering	Dbscan clustering of sets of points, for example locations of network infrastructure	Network
Area based clustering	Area based clustering of sets of points, for example locations of network infrastructure	Network
Call days	Counts of the number of days a subscriber interacted using a piece of network infrastructure	Location, Mobility
Balance of call/text events per contact	Proportion of a subscriber's interactions which was each of their contacts	Social networks, Subscriber profiling
Multi-day trajectories	Time series of a subscriber's location history	Mobility, Location
Displacement from home	Distance a subscriber is displaced in a time period from their modal location in another time period	Mobility
First seen at location	Time a subscriber made their first interaction in one of a specified set of locations	Mobility
Per subscriber daily location counts	Number of locations a subscriber has visited	Mobility
New subscribers in a time period	Number of subscribers who have not been in observed in one time period that appear in another	Network
Percentage of nocturnal calls	Proportion of a subscriber's calls which take place during the night	Poverty mapping, Subscriber profiling
Pareto interactions	Proportion of subscriber's contacts who account for some proportion of their interactions	Social networks
Proportion of interactions initiated per subscriber	Proportion of interactions with other subscribers which were initiated by that subscriber	Social networks
Radius of gyration	How widely a subscriber moves over a time period	Mobility, Poverty mapping
Aggregates of subscriber call durations	Statistics about the duration of a subscriber's calls.	Subscriber profiling, Poverty mapping
Aggregates of subscriber call durations per location	Statistics about the duration of a subscriber's calls broken down by their location when the call began.	Location, Subscriber profiling, Poverty mapping
Degree of subscriber's contact network	The number of other subscribers a subscriber interacts with.	Social networks
In degree of subscriber's contact network	The number of other subscribers a subscriber receives interactions from.	Social networks
Out degree of subscriber's contact network	The number of other subscribers a subscriber initiates interaction to.	Social networks
Subscriber Type Allocation Codes (TACs)	The type allocation codes used by a subscriber in a time period.	Poverty mapping, Subscriber profiling
Modal subscriber TAC	The type allocation code a subscriber used most often in a time period.	Poverty mapping, Subscriber profiling
Last subscriber TAC	The type allocation code a subscriber last used in a time period.	Poverty mapping, Subscriber profiling

Metric	Description	Insight categories
Subscriber handsets	The mobile phone handset models a subscriber used in a time period	Poverty mapping, Subscriber profiling
Modal subscriber handset	The mobile phone handset model a subscriber used most often in a time period	Poverty mapping, Subscriber profiling
Last subscriber TAC	The type allocation code a subscriber last used in a time period.	Poverty mapping, Subscriber profiling
Subscriber handsets	The mobile phone handset models a subscriber used in a time period	Poverty mapping, Subscriber profiling
Modal subscriber handset	The mobile phone handset model a subscriber used most often in a time period	Poverty mapping, Subscriber profiling
Last subscriber handset	The last mobile phone handset model a subscriber used in a time period	Poverty mapping, Subscriber profiling
Subscriber handset type	The type (smart, feature or basic) of mobile phone handset a subscriber used most often, or last within a time period	Poverty mapping, Subscriber profiling
Total active periods	Over a time period, how many sub periods did a subscriber make some interaction in?	Subscriber profiling, Poverty mapping
Total activity per subscriber	Total number of interactions a subscriber had in a time period	Subscriber profiling, Poverty mapping
Unique locations per subscriber	The unique set of locations a subscriber has observed interacting in over a time period	Subscriber profiling, Mobility, Location
Population weighted opportunities model	Predictive model of population movement between locations given distance and relative population of subscribers	Mobility
Louvain model	Social network community detection algorithm	Social network
Arbitrary table	Read access to arbitrary table within FlowDB, for building new metrics	Methods development
Arbitrary sql	Run access for arbitrary SQL queries to aid in prototyping new methods	Methods development
Arbitrary geographic table	Read access to arbitrary table within FlowDB containing geographic information, for building new metrics	Methods development
Random sampling	Repeatable random sample drawn from any other metric	Utility, Methods development
Subsetting	Subset by value or numeric range, of any other metric	Utility, Methods development
Arithmetic for origin destination matrices	Arithmetic operations for OD matrices, used to calculate average matrices over long time periods	Utility
Spatial aggregation	Aggregation to location of subscriber-centric metrics	Utility
Unique subscribers	All subscribers who had an interaction in a time period	Descriptives, Utility
Subsetting by groups of subscribers	Calculation of any metric using only a limited set of subscribers, specified by a subset of another metric or arbitrarily	Utility
Joins	Joins between any pair of metrics	Utility, Methods development

FLOWMINDER.ORG flowminder.org

dial Digital Impact Alliance digitalimpactalliance.org