Best Practices and Methodology for OD Matrix Creation from CDR data

Prepared by:

Dr James Goulding

N/LAB, University of Nottingham.

Tel: +44 (0) 7730 559203

Email: james.goulding@nottingham.ac.uk

Executive Summary 3

2. Introduction to CDR Data 4

3. Opportunities and Limitations of CDR data for OD matrix creation 5

4. Overview of OD matrix creation from CDR data 8

4.1 Comparison to Traditional Approaches 9

4.2 Summary of OD Matrix generation process 9

5. Repurposing CDR Data for OD-Matrices:
Theoretical challenges and solutions 10

5.1 Population Level-data Completeness and Scaling 11

5.2 Individual-level Data Completeness and Hidden Movement 14

5.2.1 Sparse Temporal Frequency Challenges 14

5.2.2 Update Frequency Bias 16

5.2.3 Bias due to Heterogeneity In Call Rates 17

5.2.4 Single Network Activity Issues 18

5.3 Location Precision Issues and False Movement 19

5.3.1 Limited precision 20

5.3.2 Non-Uniform BTS Density 21

5.3.3 Changes in BTS Operation 22

5.3.4 BTS service region overlap 23

5.4 Lack of directly recorded movement metadata 24

5.5 Privacy and Ethical Implications of Using Call Detail Records 25

6. An Agenda for Future Research 26

ANNEX 1: Physical Implementation of Origin-Destination Matrix Generation from CDR data 28

ANNEX 2: Additional Map Generation from CDR data 34

1. Executive Summary

This section considers the potential for development of micro-indicators of economic development as derived from Call Data Records (CDRs), detailing proposed methods for their creation and validation. This is joined by a critical assessment of the state-of-the-art in OD Matrix creation, the theoretical underpinnings of that process and the challenges that emerge. Finally we present an agenda for future research and development that may serve as foundation for overcoming challenges in using CDR data for mobility analysis, and to fully realise the extensive opportunities it opens up.

The exemplar data underpinning this report comes from a representative sample of approximately 10% (512,039 individuals) of anonymized mobile phone users active in the Dar es Salaam region, shared for purposes of mobility analysis by a major Tanzanian network operator for the year 2014. Section 2 introduces the nature of CDR, with Section 3 going on to discuss the opportunities and limitations of using such datasets for transport planning. Novel data streams such as CDRs allows for unprecedented insights at a scale otherwise impossible using traditional mobility analysis techniques such as Road Side Interviews. However, this potential is set against two key challenges - those of a technical nature and the absence of clear ethical and regulatory frameworks at National and International levels.

To contextualize those opportunities and challenges, Section 4 introduces the general methodology behind creation of OD Matrices from CDR data. Section 5 then provides a detailed breakdown of the theoretical and practical issues that must be addressed. Each of these limitations affects a different part of the data collection and analysis life cycle. Some of these shortcomings can be alleviated during the pre-processing stage, while others are remedied during subsequent analysis. These are drawn together into best practices of using CDR for the creation of OD Matrices, with an emphasis on how coordination with traditional techniques is key to providing optimal solutions in the future.

Section 6 extends these conclusions, providing a future research agenda that lays the foundation for overcoming identified limitations in the use of CDR data for the assessment of mobility.

Finally, in Annex 1, we provide a full, detailed breakdown of the technical process of generating OD Matrices via CDRs, along with aggregated exemplar results (which preserve both individual and commercial privacy). Here we also discuss techniques that are able to produce mobility maps, population/activity maps and sub-journey summaries. Additionally, we document the process of producing a web interface to outputs. Code snippets and further technical documentation found on the DECS OD Matrix Github repository[1] (for more extensive results and output maps please see the *Mobility Report* document).

[1] https://github.com/DECS-UK/OD-Matrices

2. Introduction to CDR Data

Call Detail Records (CDRs) are metadata (data about data) that capture subscribers' use of their cell-phones — including a timestamp, subscriber identification code and, at a minimum, the location of the phone tower that routed the call for both caller and receiver. Large operators standardly collect over six billion CDRs per day. Note that the scope of the term CDR has expanded extensively beyond the meaning implied by its original acronym ("Call Detail Records"). In current usage CDR data refers not only to calls, but to all network events (made by cell phones, tablets, etc.) which the operator records. This includes, but is not limited to, calls, sms, data usage sessions and mobile money transactions[2]. The exact form of a single CDR is dependent on the type of network event and data retention policies within the operator.

At a basic level, the CDR data that underpin OD matrix generation are expected to consist of: a timestamp corresponding to when the event took place; an anonymized identifier representing the individual initiating the event; and a unique identifier for the cell tower providing the network service to that initiator. This cell tower is formally referred to as a *Base Transceiver Station* (BTS), which is designated a unique identifier and has an associated physical location recorded. An example of such a record is shown in Table 1.

Table 1: Structure of data common across all network event types for Call Detail Records (CDRs). In this case the record is limited to a timestamp, an anonymised identifier for the event initiator (i.e. caller) and a cell coverage area identifier.

Initiating subscriber ID	Unique BTS ID	Timestamp
jggsmi13227abc	*12038097523*	*14-03-2014 00:01:12*

When the network event *type* involves two individuals (i.e. is not a data session) then the CDR will additionally contain a further two identifiers: one for the receiving individual and one for the BTS which provides the service to the receiver. Finally, the record will typically be accompanied by additional fields specific to the type of network event. Examples include call duration, sms length, data session duration and mobile money spend. Importantly, a CDR record only contains metadata about the network event but *never any of the content transmitted as part of the exchange.*

CDR's are automatically generated by network operators for billing, network management and maintenance purposes (allowing monitoring of network usage and performance[3]). However, significant potential also exists in the repurposing of these records for use in other application areas. This report focuses on one such use case - repurposing anonymized and aggregated records to understand the mobility patterns occurring across a city at a fine grain level of detail

In actuality, some form of location management of handsets by network operators is standard practice if they are to provide optimal service provision. Such tracking allows operators to direct

[2] Bias correction for these data streams is subsequently discussed in Section 5 (in theory) and in Annex 1 (in practice)

[3] D. Maldeniya, S. Lokanathan, and A. Kuramage. Origin-destination matrix estimation for sri lanka using mobile network big data. In Proc. of the13th International Conference on Social Implications of Computers in Developing Countries, 1–10, 2015.

incoming calls to the appropriate BTS in the network with optimal speed. Location management is generally undertaken by one of the following three policies:

- **Never-update**:
  Location information is not collected passively but rather in a just-in-time fashion, with all cells being 'pinged' to find out the appropriate cell to direct an incoming call to;
- **Always-update**:
  The handset informs the network whenever it is moving into a new cell. While there is no paging cost herein, networks can get quickly overwhelmed by the frequent updates;
- **Location-area-update**:
  This approach is a combination of the previous two, with BTS grouped together into Location Areas (LA). The operator is informed when a user moves to a different LA and the cells in a subscriber's LA are pinged when an incoming call is directed.

In the United States and the European Union, choice of which policy to utilize is also influenced by law that mandates operators keep track of handsets so as to provide emergency services with location approximations in emergency scenarios. Under enhanced 911 in the US[4] and enhanced 112 in the EU[5] operators have to locate users within a 50m radius in 67% of cases and 150 meters in 95% of cases. As the imposed accuracies are not achievable through BTS-only positioning, new techniques for tracking handsets have been developed. Operators can use network-centric cellphone positioning[6] or device-centric cellphone positioning[7] to effectively triangulate handsets. With never-update the prior can be prohibitively expensive to network operators in emerging economies due to the higher load imposed on the network, while with always-update the latter is restricted to smartphones with GPS capabilities.

The CDR datasets that underpin this report are not part of this mechanism, and fall under the Never-update policy remit. Thus, they only indicate a subscriber's location when that subscriber initiates or receives a network event. While data generated through more advanced tracking techniques described above could potentially improve the accuracy of insights generated via CDRs, they are unlikely to be available in emerging economies (due to both privacy issues and the prohibitive costs involved). Thankfully this lower fidelity, specifically the increase in sparsity due to the never-update policy, is of less concern due to the mass expansion of mobile phone usage in regions such as East Africa over the last decade. Network events logged within CDRs now occur at such scale and frequency across all demographics that their analysis alone is generally sufficient to generate detailed patterns of movement and mobility across a population. CDRs, however, should not be considered as a complete replacement for traditional approaches, with CDR data augmenting and significantly minimising the effort required to be spent on traditional approaches. The requirement for traditional approaches in parallel to CDR analysis is discussed in detail within this report, particularly with regard to bias correction and the inference of modality and trip purpose.

[4] J. Spinney. Mobile positioning and lbs applications.Geography, 88:256–265, 2003.

[5] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: Using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design, 33:727–748, 2006.

[6] Existing network capabilities are used to triangulate handsets based on time, angle and distance measurements generated from signal strength and handover information

[7] Where measurements and calculations are done locally within the handset

3. Opportunities and Limitations of CDR data for OD matrix creation

The use of CDR data for mobility analysis is of increasing interest to the transport planning community. Traditionally, OD matrices and associated transportation insights have been generated from Road Side Interviews (RSI). Due to logistics and costs, however, most manual counts involve a relatively small sample size occurring at any given location over a time period of a few hours[8]. Transport surveys, which have also proven an effective tool to collect data on travel trends, are equally costly to undertake, can be quickly outdated and potentially susceptible to observer biases and reporting errors[9,10,11]. Survey based methods of this nature represent significantly human intensive and logistically challenging operations.

Recently such endeavours have been augmented through data mining efforts within the public transport system (i.e. smart card access and ticketing to entrance and exit points) and more broadly using roadside cameras and Automatic Number Plate Recognition (ANPR). These approaches aim to increase the sample size (coverage) of the data collected. However, they are limited to a subset of transportation type in the former[12] case and constrained by the feasibility to deploy and maintain a city-wide camera based infrastructure in the latter.

The potential to use CDR data to underpin mobility analysis significantly advances the trend away from time consuming and/or resource intensive techniques. Repurposing CDR data provides several advantages: unparalleled **scale** (in terms of sample size); **coverage** (in terms of observed spatial area and modes of transportation); **spatial granularity** (in terms of the precision of origin/destination units that can be outputs) and **temporal fidelity**[13]. This occurs at orders of magnitude reduced infrastructural and human resource costs.

Generation of OD Matrices from this relatively new form of data naturally has both strengths and weaknesses in comparison to conventional approaches. The coverage, precision and eliminated collection cost that CDR analysis promises must be considered in the light of limitations on what can be derived from such data. Data that has not been collected for the sole purpose of transport transport planning increases the risk of a range of **biases** that may occur within the data - biases that need careful assessment and adjusting for. Identification of **short trips** also becomes difficult, as does identification of **mode of transport** when they cannot be directly surveyed. Such problems - inherent to passive data collection - stand in in contrast to the active nature of surveys and other traditional approaches, which allow one to directly query transport motivations (such as understanding mode-choice).

[8] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. A system for real-time monitoring of urbanmobility using cell phones: A case study in rome.IEEE Transactions on Intelligent TransportationSystems, 12(1):141–151, 2011.

[9] R. M. Groves. Nonresponse rates and non-response bias in household surveys.Public Opinion Quarterly,70(5):646–675, 200.

[10] J. C. Herrera, S. Amin, A. Bayen, S. Madanat,M. Zhang, Y. Nie, Z. Qian, Y. Lou, Y. Yin, and M. Li.Dynamic estimation of od matrices for freeways and arterials. Technical report, Institute of Transportation Studies, 2007.

[11] J. C. Herrera, D. B. Work, R. Herring, X. Ban,Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: the mobile century field experiment.Transportation Research Part C: Emerging Technologies,18(4):568–583, 2010.

[12] I.e. formal public transport systems where smart cards are in operation. In Dar es Salaam, for example, this would not include the extensively used bus (Dala Dala) network which operates on a semi-formal basis and loosely regulated basis

[13] The impact of these notions is discussed in much further detail in Section 5

These challenges evidence the fact that, while there are immense benefits of using CDR data at such scale, there is still a valuable place for traditional approaches in the analysis process. Indeed, with CDR-driven matrices providing a cornerstone for mobility analysis, traditional techniques can then be used to augment results with increased parsimony and precision, fine tuning and ground truthing OD matrices. In order to illustrate the advantages and characteristics of both CDR and traditional Road Side Interview techniques, a comparison is provided in Table 2.

In the generation of OD Matrices for the Dar es Salaam region of Tanzania we have also identified specific technical challenges that must be carefully accounted for in the use of CDR data in order to realize their potential. These challenges and associated solutions are enumerated in depth in section 4, with Annex 1 documenting a practical and technical description of how these solutions were implemented.

Table 2: Comparison of Call Data Record and Roadside Interview Data for supporting the generation of Origin-Destination Matrices, adapted from Tolouei et al. (2015)[14] and extended via the analysis described in section 4 onwards.

Issue	Roadside Interview (RSI) Data	Call Detail Record (CDR) Data
Type of data source	Cross-sectional *(sample for a single period, normally days)*	Longitudinal *(cross-sectional data collected over an extended period, normally a year)*
Collection Approach	Active *(for purposes of Transport Planning)*	Passive (for network operation purposes)
Sampling Rate	Very Low *(for any given road 10-20% of vehicles, with the number of roads sampled limited by available resources and time-scales)*	High *(for the whole city, the percentage of population covered by network provider. This sample size is vast compared to other technologies)*
Processing Costs	High *(RSI are a significantly labour-intensive process, and data collection alone can take months to coordinate successfully)*	Low *(Once processes are in place, analysis of passive CDR data can be performed in a few days using appropriate Big Data technologies)*
Variation Observed	Spatial Variation of Trips *(Assumes stationarity in behaviour)*	Spatial and Temporal Variation of Trips *(Allows for non-stationarity in behaviour)*
Data Bias	Potential for Response Bias *(minimizable via careful survey design and sampling strategy)*	Potential for Bias towards the profile of subscribers compared to the full population *(minimizable via careful analysis of the demographics of the subscriber base)*
Extrapolation of Sample	Independent and well understood *(combining count data with statistical modelling where journeys traverse more than one sample site)*	Dependent and under continued development *(requires ground truthing data on how mobile phone user's behaviour is representative of the whole population)*
Coverage of Purpose	Limited *(Focussed to main roads and long trips)*	Very High *(All purposes apart from micro-trips covered)*
Identification of Purpose	Straightforward *(Via a survey question in the interview)*	Complex *(Requires inferring via modelling assumptions or rule-based domain knowledge)*
Coverage of Mode	Limited	Very High

[14] SR Tolouei, P. Alvarez, and N. Duduta. Developing and Verifying Origin-Destination Matrices using Mobile Phone Data: The LLITM Case, Proceedings of the 2015 European Transport Conference.

	(Focused on Private Motorized Transport)	*(Covers all modes of transport used)*
Identification of Mode	Straightforward (Via a survey question in the interview)	Complex *(Requires inferring via modelling assumptions or rule-based domain knowledge)*
Geographical Scope	Limited *(Only movement intercepted by screenlines / cordons can be assessed)*	Extremely Wide *(In theory all movements outside of short trips occurring within a single BTS catchment area can be assessed)*

4. Overview of OD matrix creation from CDR data

We now move on to the process of generating OD matrices from raw CDR data. At its heart, this task requires us to process the mass set of digital traces left behind by the population as they traverse their daily lives. From those traces we must attempt to reconstruct the journeys that they undertook over that period. While the number of people represented within the dataset is immense, each individual is only represented by a sparse sample of the locations they actually visited[15]. Therefore, technically we must transform a vast number of sparsely sampled event series (i.e. expressing each individual's location history only when a network event occurs) into a mass set of 'journey' constructs, all of which must meet pre-defined validity conditions. This set can then be aggregated, scaled and interpolated to produce our final OD matrices.

Unlike traditional analyses, the nature of this data mining approach forces us to first provide rigid, formal definitions of exactly what we mean by the terms *origin*, *destination* and *journey*. Here, both origins and destinations are subcategories of the overarching concept of a '*stop*'. A stop is defined as a set of contiguous network events that occur at the same location, over a minimum period of time. This notion is parameterized to ensure we have sufficient confidence that any stop we have detected is not a transient location, but actually a location that the individual has actually *settled* in.

An algorithm must consequently be developed to exhaustively mine each person's event series for such stops. Once achieved, the algorithm must next detect pairs of contiguous stops which occur at different locations and hence reveal movement. This pair can then be designated as a *journey* - the initiating stop becoming the journey's *origin*, and the concluding stop as the journey's *destination*. A further filtering algorithm can then be utilized to process the resulting journeys in order to refine results and ensure that each journey fulfills certain strict conditions (ensuring, for example, that we have sufficient confidence that the journey did not go via any *other* stops and hence represents a true direct trip, or ensuring that the generating individual was not exhibiting spurious or outlying behaviour - such as might be exhibited by the traces of bus drivers)[16].

Resultant journey data is then transformed (via aggregation and counting) into an intermediary Origin Destination matrix. This intermediary only reflects the subset of journey's that we have been able to identify in our sample population (specific users of that network operator. Consequently, this matrix must still be scaled to estimate the behaviour of the population as whole. The scaling process is performed as follows: first, the journey set is processed so as to produce a model of all movements occurring across the region at specific time periods. This allows computation of traffic counts at a set of key locations at those time periods. A scaling factor can then be calculated so as

[15] Discussed in detail in Section 5.2
[16] Discussed in detail in Section 5.3

to minimize the error between these *computed counts* and *ground-truth* counts directly observed at those test locations.

Once scaling has been achieved, interpolation to a desired zonal representation can finally be undertaken, further protecting both individual and commercial privacy[17] - and producing our final output OD matrices ready for practical usage in transport planning.

4.1 Comparison to Traditional Approaches

Prior to progressing to best practices within this process, it is important to consider the key differences between a CDR based approach compared to its traditional RSI based counterpart. As detailed above, CDR-based OD matrix creation requires the reconstruction of a population's journeys (and associated metadata) from 1. A sparse sample of their location history[18]; and 2. Any available external information such as transport infrastructure and ground-truthing counts.

This is in contrast to OD matrix creation from RSI which focuses on an extremely small sample of the population - but which is able to obtain fully declared journeys and metadata. For CDR such metadata has to be inferred. Traditional approaches also relinquish necessity for strict definitions of concepts such as *origin, destination* and *journeys (which are inherently nebulous)*. When OD matrices are formulated from roadside interviews, such concepts may be left somewhat subjective and/or based on criteria from the individual's broad context. Nevertheless, in both cases the resultant OD matrices indicate movement from one location to another.

4.2 Summary of OD Matrix generation process

Based on a fixed definition of origins, destinations and journeys, we can now enumerate the high level process that must be undertaken when constructing OD matrices from CDR data:

1. **Data Cleansing:**
   Raw data must be converted into network event series. At this point both any BTS and spatial regions can be merged if appropriate (e.g. to deal with intermittent towers, or towers located in such close proximity that the location of user cannot be distinguished);
2. **Stop Identification:**
   A series of potential *origins* and *destinations* much then be extracted from each event stream. These are typically mined under a combined definition as *stops* at the individual level, with stop's being sub-categorized as *transient* or *semi-permanent*. Stops are parameterized by the number of contiguous network events required in the stop, the minimum duration time as a whole, and the maximum inter-event time;
3. **Journey Generation:**
   For each individual, their resulting series of stops must then be converted into <origin,destination> journey pairs. At this point an appropriate score for confidence in the efficacy of the pair as a direct journey can also be attributed to it;
4. **Journey Cleansing:**
   Any individuals with outlying behaviour must then be removed to reduce bias in the dataset. Journey's which do not meet some predetermined confidence threshold may also be removed at this point;

[17] For more information see Section 5.5 *Privacy and Ethical Implications of Using Call Detail Records*

[18] Sparsity here refers to the sparsity of location samples observed for each individual (for further discussion see Section 5). This is distinct from the sampling of the individuals themselves from the overall population (the challenges of which are discussed in section 5.1).

5. **Metadata Tagging:**
   Surviving journeys may then be algorithmically tagged with metadata (such as the journey's purpose or mode of transport) via further inference. This inference is based on further processing of the CDR data or integration of external domain knowledge;
6. **Intermediary OD Matrix Generation:**
   The resultant journey set can then be filtered dependent on the task at hand before aggregating and counting to produce an appropriate OD matrix representation. This filtering may be temporal (e.g. weekends), spatial (specific wards) or via metadata categories (such as commutes);
7. **Scaling:**
   Raw OD Matrices must then be scaled in order to correctly extrapolate the number of detected journeys to the full population. By necessity this process requires some external ground-truthing data for validation;
8. **Final OD Matrix Generation**:
   Finally, results can be interpolated into a specific geospatial representation. This serves to protect both individual and commercial privacy, with matrices being projected to a disjoint set of geographical regions *(zonal units)* that underpin the final output OD matrix and match the task at hand.

While this process is conceptually straightforward and produces OD matrices of unprecedented scale and fidelity, the repurposed nature of CDR data means that its properties can present a range of technical challenges in many of these steps. These technical challenges and their respective solutions, along with definitions of origins, destinations and journeys they admit are discussed in Section 5. Full details of how theoretical solutions are implemented in our final approach are provided in Annex 1.

5. Repurposing CDR Data for OD-Matrices: *Theoretical challenges and solutions*

CDR data contains significant information from which to build high fidelity origin destination matrices at a previously unseen scale. Not designed for this purpose, however, the properties of the CDR data (i.e. data is only recorded about subscriber events rather than movement directly) are not optimal and require careful processing. In addition, as data is pertaining to individuals, there holds the potential for raising privacy concerns if not handled correctly. Below we identify the challenges which face OD matrix generation via CDR data (which can broadly be grouped into five categories), and then proceed to break them down in more detail along with best practice solutions in the following subsections.

Challenges of using CDR Data for OD Matrix Generation	
5.1	**Population-level data completeness and scaling** CDR data is restricted to both the subset of the population which utilize cellular devices and those that choose the network provider providing the CDR data. This introduces potential biases in the population being modeled, and scaling issues in understanding how to accurately extend to represent the whole population.

5.2 **Individual-level data completeness and hidden movement**

CDR data only contains events when an individual initiates or receives a network event. As such movement is not directly observed, and must be inferred from individual event streams - which are themselves *sparse* (depending on network usage patterns and loyalty). This introduces technical challenges pertaining to journey reconstruction and how to handle issues of "hidden movement" in individual streams.

5.3 **Location precision and false movement**

CDR data does not record individual's locations exactly, but rather as a non-precise proxy via the location of the infrastructure delivering the service. This introduces various challenges including: 1. How to handle the limited precision afforded by CDR (in comparison to positioning technologies such as GPS); 2. How to integrate different granularities occurring at geographical level, occurring due to varying local infrastructure density; 3. How to accommodate for precision that varies temporally due to infrastructure issues/upgrades; and 4. How to handle issues relating to overlapping service areas, handovers and load balancing mechanisms (which can produce "false movement").

5.4 **Lack of directly recorded movement metadata**

In contrast to traditional surveys or RSIs, repurposed CDR data cannot directly capture many of the movement metadata, such as transport mode, route and/or motivation for undertaking a journey. To solve this, methods are required for computationally inferring and verifying those metadata, strategically augmenting this analysis utilizing alternative data sources and more traditional interview and surveys.

5.5 **Privacy**

Due to the private nature of the data encoded in CDR's such repurposing must only occur on data where user identifiers have been pre-anonymised by the network provider as recommended by the Groupe Special Mobile (GSMA), the mobile telecom industry body, in their emergent guidelines on the use of CDR data[19]. This is a necessary, but not sufficient, condition to addressing privacy concerns. Specifically, additional steps need to be undertaken to prevent the re-identification of individuals through individual specific patterns within the data and external information. This is discussed, along with the solution undertaken in this work, in section 5.5 (along with a practical implementation of the solution in Annex 1).

5.1 Population Level-data Completeness and Scaling

Challenge: CDR data contains records for all subscribers of the network operator providing the data, this however is is not a complete sample of the population. This is caused by competing operators and lower than 100% population uptake of cellular services. This results in reduced datasets with respect to the population. Moreover, each CDR dataset is likely to be biased towards some sector of the population. Operators often to market to different population groups, and those without cellular service often fall into lower socio-economic and/or age group demographics. This means care must be taken not to bias results towards the specific demographic of users

[19] GSMA. 2014. GSMA Guidelines on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak. Retrieved March 20, 2016 from http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak-_October-2014.pdf

subscribing to the network provider's services. As a consequence all processing results must be considered as intermediary in their raw form, offering relative, rather than absolute, origin-destination counts until scaling occurs.

Discussion & Solution: If no systematic bias exists in then scaling to the whole population can be approached by a range of measures which are discussed below (although, as will be expanded upon, we will ultimately recommend the use of ground truth camera-based measurements in conjunction with known statistics on modality ratios to determine the correct scaling factor for accurate journey count estimates for the whole population).

Research in the area of OD matrix scaling has focused on either re-purposing existing statistics (e.g. census figures) or on the strategic sampling of ground truths (at key transport locations) fusing the more traditional approaches such as road-side interviews and traffic counting into the process. One data-driven approach to scaling that has been explored in the literature is to estimate a count for residents in each region from the CDR data, and to compute a scaling factor that matches actual population sizes in those regions (commonly referred to as expansion factor) against known population statistics[20]. This scaling factor is then used as a surrogate to scale the OD matrices counts.

While such an approach may be appealing if up-to-date population statistics exist, this is rarely the case in developing countries (and particularly not at the spatial/temporal granularity we desire). The use of this sort of surrogate scaling factor of could then actually introduces a risk of further error - the assumption that resident counts estimated from CDR data will reflect true population levels is clouded by issues of cell phone demographic usage[21,22].

Another alternative approach has used *mode choice probability*, *vehicle occupancy* and *usage ratios* to determine scaling factors. However, a heavy reliance on the availability of accurate traffic and demographic data in a high spatial resolution remains[23]. Equivalently, cellular penetration, mobile phone non-usage and vehicle usage have been used in an attempt to calibrate scale factors in computer simulations. While this approach to scaling has shown promise when used in a micro-simulation for Dhaka[24], transferability from the microsimulation model of traffic points to real-world scenarios in other study areas has not been tested in detail. Driver decisions which are modelled in such simulations in particular may be less accurate in rapidly urbanizing cities such as Dar es Salaam.

As a **best practice**, we therefore recommend the identification and ground truth sampling of multiple important location points over the region, examining both different times of day and repeating on multiple days. This approach enables not only a scaling factor to be computed, but additionally provides an opportunity for confidence measures to be developed (Figure 1 shows the points taken

[20] S. Colak, L.P. Alexander, B.G. Alvim, S.R. Mehndiratta, and M.C. Gonzalez. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Record: Journal of the Transport Research Board, 2526: 126-135, 2015.
[21] L. Alexander, S. Jiang, M. Murga, and M.C. Gonzalez. Origin Destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies, 58:part B,, 240-250, 2015.
[22] M. G. Demisse, et al. Inferring origin-destination flows using mobile phone data: A case study of senegal. In Electrical Engineering / Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016 13th International Conference on , 2016.
[23] P. Wang, et al.. Understanding Road Usage Patterns in Urban Areas. Scientific Reports, 2, 1-6. 2012.
[24] Md. S. Iqbal et al. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C 2014.

in Dar es Salaam). This is achieved by computing scaling factors for each ground truth sample point and considering their variance. In the practical implementation detailed in Annex 1, this approach was undertaken using video camera based traffic counting at multiple traffic sample points. Note that such an approach leads to OD matrices for vehicular journeys counts, with external modal statistics required and/or more complex surveying required to extrapolate to full modality OD matrices. There is *much potential* to automate this work via computer-vision and deep learning algorithms. When combined with deployment of a small number of fixed cameras would provide sophisticated scaling factors which vary over time and geographical region (see section 6: "An agenda for future research").

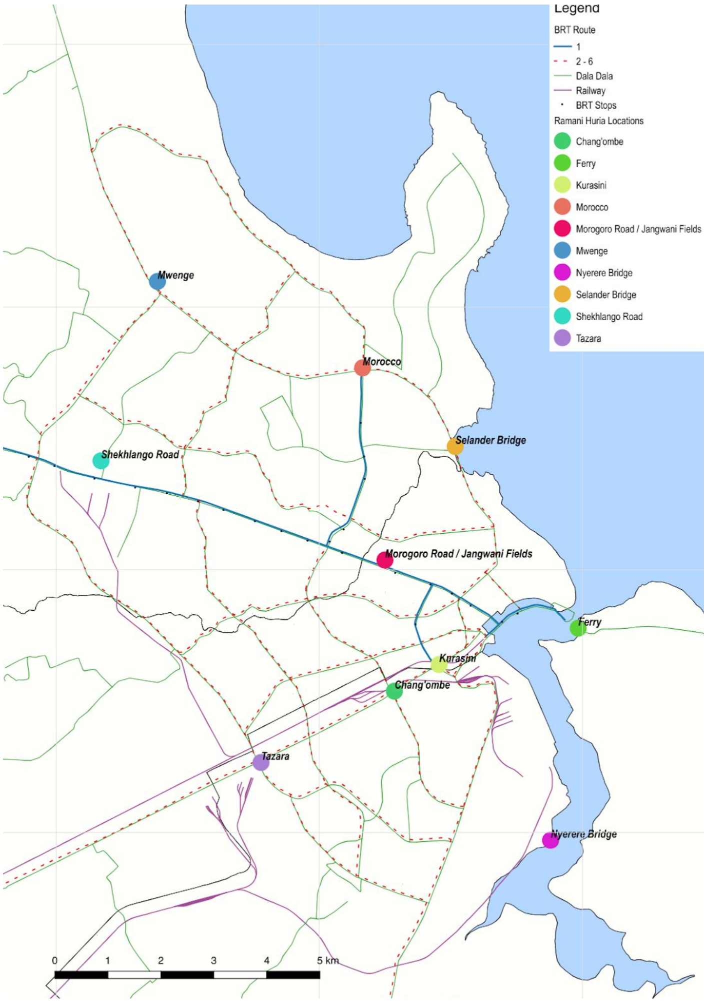


Figure 1: Example Identification of key flow points within the Dar es Salaam Region. In order to perform scaling, ground truth data was obtaining at each of these points, with cameras collecting footage across multiple days at three daily time periods (morning, day and evening). Traffic counts could then be produced and cross-referenced with model estimates to produce a set of robust scaling factors.

Unlike scaling factors, issues of population bias can be complex to accurately resolve. In urban spaces in Tanzania, where mobile phone penetration has been reported to be close to 92%[25,26], bias in uptake may (relatively safely) be assumed to be negligible and left uncorrected. Operator bias is then required to be assessed based on the known coverage of the operator in the area of interest and a qualitative assessment to the viability of using the operator's CDR undertaken. Alternatively OD matrices could be computed assuming no systematic bias and a confidence measure for the

[25] Ministry of Health, Community Development, Gender, Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, and ICF. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. 2016, Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF. Retrieved December 18, 2016 from http://dhsprogram.com/pubs/pdf/FR321/FR321.pdf.

[26] Compared to a 79% penetration rate in Tanzania overall (Tanzania Communications Regulatory Authority. Quarterly Communications Statistics Report. April-June 2016 Quarter. (https://www.tcra.go.tz/images/documents/telecommunication/CommStatJune16.pdf)

scaling factor computed as discussed above. Low confidence in such a measure indicates error within the system potentially indicating the presence of such bias. **Best practice** must therefore currently focus on the acquisition of a provider with sufficiently unbiased population coverage (until improved geo-demographic intelligence can be sourced in emerging economies). This must be coupled with comprehensive measurement of scaling factor error to ensure fit-for-purpose and confidence. Initial studies have shown this approach to be successful with, for instance, mobility models in Kenya that have been developed from CDR data producing similar insights as manually constructed ground truth data[27].

5.2 Individual-level Data Completeness and Hidden Movement

In an ideal world, all of a population's movement would be available for aggregation into a mobility analysis. However, this is not the case in CDR data, which only record an individual's location when a network events occur (whether calls, sms, data sessions, mobile money transactions). This results in a sparse stream of events, containing a subset of each individual's actual behaviour. This in turn leads to several issues that must be ameliorated, of which four of the most pressing will be broken down as follows:

- ❏ *Section 5.2.1: "Sparse Temporal Frequency Challenges"* considers how to handle the issues which arise around sparsity in general, most central of which is "hidden movement";
- ❏ *Section 5.2.2: "Update Frequency Bias"* addresses the fact that some individuals' event series will also be far more sparse than others, introducing the danger that OD matrices will be biased towards those people with high usage patterns;
- ❏ *Section 5.2.3: "Bias due to Heterogeneity in Call Rates"* considers that activity in certain regions may also introduce biases - areas with forced waiting time, for example, (i.e. at ferry terminals) may result in overestimations of stop frequencies, and this needs accounting for;
- ❏ *Section 5.2.4: "Single Network Activity Issues"* considers that subscribers may have multiple SIM cards acting over multiple networks, resulting in a single network provider's CDR data only containing a biased lens into an individual's movement.

5.2.1 Sparse Temporal Frequency Challenges

Challenge: Observing movement accurately requires the continuous, precise sampling of an individual's location (e.g. as within sensors such as GPS) at low temporal intervals. In contrast, CDR data is only recorded when a network event takes place. Depending on an individual's usage pattern the time between these events could range from a few seconds to multiple hours, resulting in significant amounts of *unobserved* movement (often referred to as "hidden movement"). As such, two sequential records involving a change in location may involve any number of unobserved sub-journeys and/or periods of non-movement in between them of which we have no indication. Furthermore, a lack of samples reduces the certainty as to whether an individual has actually *stopped* in a location when we see a network event occurring at that location (i.e. are we truly observing a journey origin or destination when we examine a network event's location, or merely some intermediary point within their travels?).

[27] A. Wesolowski, N. Eagle, A.M. Noor, R.W. Snow, and C. O. Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. Interface (Journal of the Royal Society), 10(81), 2013.

Discussion & Solution: The sparsity in recorded locations in CDR data means that not all stops in an individual's daily trajectory can necessarily be identified. Thankfully, the sheer scale of the dataset render hidden movement in a particular individual less of a problem - the key is that we have enough samples across the population to fill in the gaps, and this is where CDR data is at it's strongest. In order to construct valid OD matrices, existing approaches therefore (correctly, due to sample size) focus entirely on reduction of *false positives* when detecting stops. Such an approach has the effect of actually increasing the sparsity of sequential events from which to identify journeys, but as will be discussed - this is an acceptable concession in any individual stream. Equally the sparseness of events prevents the identification of any journeys with either the origin, destination or both missing from the CDR data. Yet again ensuring that false positives (false movement) are eliminated is essential. With enough data, and under the assumption that phone usage across stops is not systematically biased at different locations (see subsection 5.2.3: *Heterogeneity in Call Rates*), then such loss in individual series is ameliorated by the sheer scale of the dataset. More pressure is necessarily then applied to determining an accurate population scaling factor (see Section 5.1), but this is entirely achievable with thorough ground truthing - and we are left with confidence that the journey's detected represent complete trips, and not sub-journeys.

An alternative solution to the issue of hidden movement is to integrate the fact that human mobility is known to be highly predictable - it is now well documented that humans tend to follow similar patterns over time, such as commuting from home to work[28,29,30]. This fact can be leveraged by viewing a set of days (say all weekdays in a month) as a repeated *sample* of behaviour over a single day and modelling behaviour statistically as a counting process. Assuming each individual follows a similar daily pattern over this period of temporal aggregation, aggregation of this type directly can significantly decrease the sparsity of the events. Such an approach will increase and affirm the presence of regular journeys at a higher granularity (in terms of sub-journey information) at the expense of uncommon ones. While this is an extremely promising solution, research into the the full effect of the modelling technique and optimal periods for aggregation is still required - and hence not undertaken in this work.

The above considerations, however, generate the following best practices. When mining for journey's the definitions provided below will provide the most robust outputs when we are availed with data at scale:

Best Practice "Journey" Definition - with respect to an OD matrix, a journey should refer to a high-confidence movement from one stop to another. The focus should be on ensuring efficacy of origins/destinations and defending against mis-casting of intermediate stops as true destinations. This policy enforces a strict answer on questions such as "if someone is travelling from one city to another, but stop for lunch, is this two trips or one?", or "Does it make a difference if this is a sandwich at a roadside petrol station while fueling the car with fuel or involves stopping in the CBD and catching up with friends?". This policy mandates that all efforts ought be made to ensure sub-journeys are *omitted* in any raw OD matrix - unless a transient matrix is actively sought.

[28] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. Science, 327(5968), 1018–1021, 2010.

[29] R. A. Becker, R. Caceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp 11), 2011.

[30] G. Smith, R. Wieser, J. Goulding, and D. Barrack. A refined limit on the predictability of human mobility. Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on. IEEE, 2014.

Setting high thresholds for acceptable journey identification means that we may, as a consequence, miss valid stops and hence. This raises the danger of mis-classifying movement as direct, when in reality it was composed of two or more individual journeys ("misinterpreted movement"). To combat this, our definition of a journey must also therefore enforce a temporal bounding. This requires a upper temporal limit, above which the likelihood we have missed a stop becomes unacceptable. This is accompanied by a lower limit, below which movement would have been unfeasibly fast. For how confidence and temporal thresholds can be physically implemented please see Annex 1.

Best Practice "Stop" Definition - commonly a stop is considered as a location at which an individual spends some minimum threshold of time and can therefore be assumed to have 'settled'. This *stop based approach*[31,32] typically defines a location either as a single BTS or within a fixed radius and/or a parameterized set of nearby towers (due to service provision overlap - see section 5.3.4). Note, that if the minimum time threshold is set to zero, OD matrices will degenerate to what is known as a *transient approach*[33,34] containing all sub-journeys - and won't truly describe movements between origins and destinations at all *per se*. Elimination of static trips/false positives is therefore the priority.

A balance is required here however - as we extend the minimum time events must span to be deemed a valid stop, we introduce the risk that an individual may have left and returned within that period ("splitting the stop"). To combat this a stop definition must also integrate the notion of a maximum *inter-event time* permissible between sequential network events at the same location. Thus, a stop is strictly defined around two thresholds: 1. the minimum amount of time at least two contiguous events occurring at the same location must span; and 2. a maximum inter-event time below which it is considered unlikely there will have been unobserved movement (i.e. where the individual will have left the location and come back). Again, physical implementations of these requirements can be seen in Annex 1.

Final Remark: While not as pronounced in emerging economies, there is an increasing shift away from SMS and phone calls toward mobile data usage. Assuming data events are logged, this transition will likely aid the generation of OD matrices, reducing the sparsity of individuals logs with the average inter-event time generally being lower compared to SMS and cellular calls[35,36], as data applications typically request data at significantly higher frequencies. Finally, if possible, movement events (between tower locations) could be logged by the data provider. While these are not recorded for billing and hence not typically available (as is the case in the dataset underpinning this work) working with the data providers to expose this information would eliminate many of the discussed challenges.

[31] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M.C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. Scientific Reports, 2, 1-6. 2012.

[32] Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C, 40, 63-74, 2014.

[33] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M.C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. Scientific Reports, 2, 1-6. 2012.

[34] Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C, 40, 63-74, 2014.

[35] F. Calabrese. The Geography of Taste: Analysing Cell-Phone Mobility and Social Events. Proceedings of the International Conference on Pervasive Computing, 22-37, 2010.

[36] C. Chen, L. Bian, and J. Mac. From traces to trajectories: How well can we guess activity locations from mobile phone traces?. Transportation Research Part C, 46, 2014.

5.2.2 Update Frequency Bias

Challenge: As CDR data is only created when network events take place, users with a high network activity (and consequently their journeys) are over-proportionally represented within the data set[37,38]. Without further adjustments, this over-representation can significantly affect the representativeness of any predictions based on CDR data alone[39].

Discussion & Solution: Two options exist to deal with the issue of *update frequency bias*. The first is the identification and removal of over-proportionally represented individuals identified as outliers[40]. Alternatively, journey contributions per individual to each OD matrix can be normalised by weighting contribution counts per individual by either 1. their total number of identified journeys or 2. the scale of their network activity. While this will artificially reduce intermediary OD matrices, this will have more impact on the distribution of journeys rather than total journey counts (due to the scaling approach implemented). More research is currently required in the field to understand the impacts of such weighting individual contributions to OD matrices, and thus **best practice** recommendation is currently to use the former solution - outlier removal (as we do in Annex 1). While there appears much promise in the (potentially superior) weighting approach further research efforts are required (as detailed in section 6: "A future research agenda").

5.2.3 Bias due to Heterogeneity In Call Rates

Challenge: The generation of network events by individuals may be either positively or negatively biased within certain spatio-temporal regions. This may artificially increase/decrease the probability of an origin or destination being identified *in comparison* to other areas. This may introduce errors when scaling factors are used to generalise the observed results to the population, over emphasizing journeys to zones where high network usage occurs. This bias can cut both ways - an example when negative bias may occur is in locations such as home or work where landlines may be used in favour of mobile connections. Conversely, an example of positive bias is within waiting areas such as bus or ferry terminals network activity where subscribers are potentially using their phones at higher rates to 'kill time'.

Discussion & Solution: Negative areal bias due to reduced mobile usage is typically considered less of a concern within emerging economies and hence **best practice** recommendation is, currently, to leave this unadjusted. This is due to the very low number of landline available in emerging economies (With only an estimated 2% penetration across Sub-Saharan Africa[41]) and 'mobile first' usage patterns. Sub-Saharan Africa mobile connections have been shown to be almost 50x greater than landline connections[42].

[37] T. Couronne, A-M. Olteanu, and Z. Smoreda.. Urban mobility: Velocity and uncertainty in mobile phone data. In Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, 1425–1430, 2011.

[38] T. Couronne, Z. Smoreda, and A.-M. Olteanu. Chatty mobiles: Individual mobility and communication patterns. In NetMob 2011, 2011.

[39] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41(22):1–11, 2008.

[40] H. Wang, F. Calabrese, G.D. Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call records. In 13th International IEEE Conference on Intelligent Transportation Systems, 318-323, 2010.

[41] Pew Research Center. Cell Phones in Africa: Communication Lifeline. 2015, April 15. Retrieved December 10, 2016 from http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/

[42] GSMA. Gauging the relationship between fixed and mobile penetration. Technical report, GSMA, 2014.

In contrast, **best practice** recommendation is to actively adjust for positive areal bias via ground truth information, identified in data sources such as public transport maps and/or ferry port and airport locations in the calculation of the scaling factor[43]. Currently these adjustments must be guided by heuristics, but there is much potential for research to automate this and generate a dynamic scaling map (varying scaling factors both geographically and temporally) across an extent. Additionally, outlying regions (in terms of exceptional activity across all individuals in the CDR data) can be identified and investigated. Adjustments could also be made to journey counts via custom scaling factors through a small number of targeted observations using more traditional approaches such as interviews, surveys or (automated) traffic counting.

5.2.4 Single Network Activity Issues

Challenge: Any CDR dataset sourced from a single operator is necessarily limited to a subset of the population (and consequently a specific demographic distribution) who have self-selected to be subscribers to that service. Additionally, with the advent of multi-sim devices, many subscribers may be using multiple operators, with each operator subsequently only receiving a sample of overall network events biased towards certain usage situations (e.g. calls on one sim, data on another). Such a situation is not uncommon to mobile users in emerging economies. Motivating factors for multi-sim use include: 1. The ability to offset some of the issues associated with limited network coverage when travelling; 2. The mitigation of network down time, which is a particular concern in rural areas within emerging economies; and 3. The reduction of costs through the use of different SIM cards for calls to different user groups[44,45] utilizing cheaper intra-network call tariffs and/or different SIMs for different network services (e.g. Data vs voice and sms). The bias introduced in assuming that mobility analysis generated from Single Network Activity is representative of overall behaviour can be broken down into three main challenges: 1. The systematic loss of network events due to usage behaviour; 2. The temporary loss (or gain) of network events due to the CDR operator's network (or their competitors) becoming temporarily unavailable; and 3. The systematic, per individual, loss of events in a given region as they change providers to maintain/improve service.

Discussion & Solution: Systematic occlusion of network events due to usage behaviour can be considered to be a manifestation of individuals with increased sparse temporal frequency bias and low update frequency bias. Correction for these biases is therefore often sufficient as described in section 5.2.1. If the occlusion is geographically systematic (i.e. individuals changing providers in a given region for personal and/or network reasons) then the resultant bias is similar to - and can hence be addressed in a similar fashion to - heterogeneity in call rates (Section 5.2.3). Specifically, areas of known reduced operator coverage, or operator penetration, can be selected for examination and manual correction. Additionally analysis of low usage areas, compared to known population statistics and network operator penetration rates can be done, again to select areas for examination and manual correction based on the deployment of small numbers of targeted more traditional observations.

[43] Md. S. Iqbal, C.F. Choudhury, P.Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C, 40, 63-74, 2014.

[44] K. Ding. The Disintegration of Production: Firm Strategy and Industrial Development in China, chapter The specialized market system: the market exploration of small businesses, pages 149–176.Edward Alger Publishing Limited, Cheltenham, UK,2014

[45] A. M. Desai and E. Forrest. E-Marketing in Developed And Developing Countries: Emerging Practices,chapter Mobile Marketing: The Imminent Predominance of the Smartphone, pages 97–115. Business Science Reference, Hershey, PA, USA, 2013

Non-systematic events with large scale impact, such as regionwide network downtime, however require specific investigation and potential manual adjustment. Due to their global and significant impact on the CDR data for a given region these are easily identified. Corrections then may include omission of the temporal/spatial period with global rescaling to the OD matrices to account for the missing data, data imputation or another bespoke adjustment depending on the severity and spatio-temporal area affected and the regularity of the problems.

Final Remarks: Many of the discussed single network activity biases are significantly more pronounced in areas where operator coverage is reduced and infrastructure less reliable. In urban areas with good coverage and reliable infrastructure adjustments may not even be required. Finally, while not discussed, these biases are obviously less of an issue in the (atypical) situation where multiple operators have contributed to the CDR dataset underpinning the mobility analysis.

5.3 Location Precision Issues and False Movement

A key consideration when working with CDR data is that it does not specifically record the location of each device that it services, but rather the base transceiver station (BTS) that provided the service. It is therefore the physical location of the BTS station that is used as a proxy for an individual's location within CDR data. Each BTS is surrounded by a region within which it is able to provide network service. Depending on the density of the BTSs these may overlap.

While not always the case, in general subscribers will be serviced by their closest (strongest signal) BTS[46]. Following discussion with network engineers and analysing with them the complexities of cell operation, **best practice** recommendation is that subscribers should be assumed to be within the "voronoi" catchment region surrounding the BTS handling their network event. A voronoi region (or cell) for a particular BTS consists of all points closer to that BTS than any other. The combination of every voronoi region is referred to as a voronoi map. An example of how such voronoi regions physically manifest is provided in Figure 2.

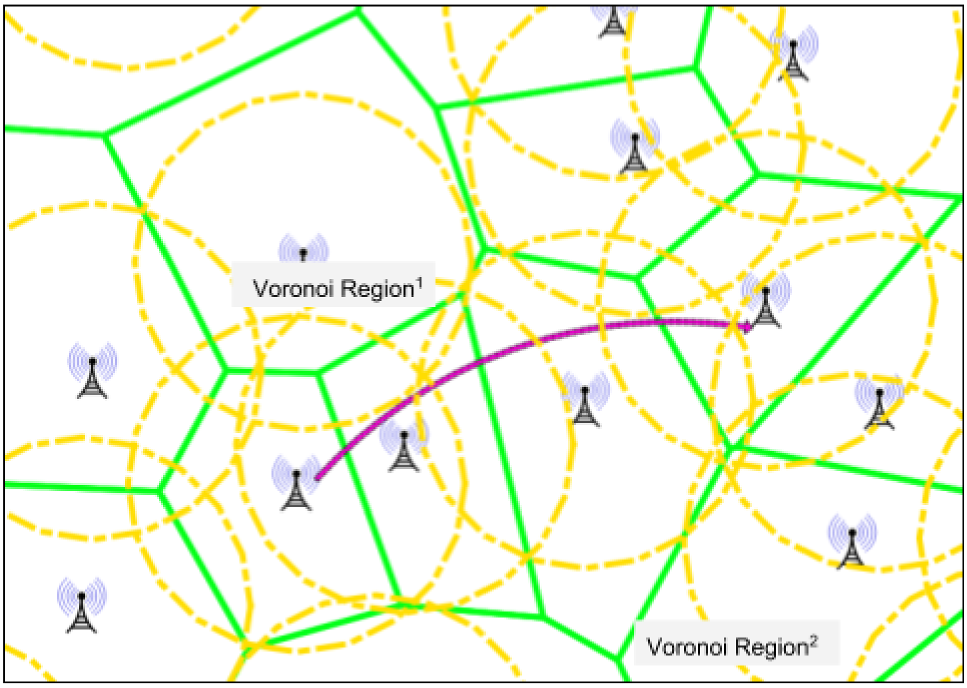


Figure 2. Potential BTS catchment area (yellow) contrasted with a Voronoi cell model, which reflects the most likely location region of a user. An example movement path of a user is illustrated in purple.

[46] P. McGuiggan. GPRS in Practice: A companion to the specifications. Chapter: Operation in the Physical Layers. Wiley

While this assumption maximises the likelihood of correctly positioning a subscriber when constructing OD matrices (but certainly doesn't guarantee a correct positioning), several challenges remain, which can be categorized as follows:

- ❏ *Section 5.3.1: "Limited Precision"* examines the issues that arise due to the non-precise, areal nature of these voronoi locations;
- ❏ *Section 5.3.2: ""Non-uniform BTS density"* addresses the fact that the size of each voronoi cell varies significantly across the extent of both city regions, and countries as a whole (depending on infrastructure density and broadcast capacity) - a factor which must be taken into consideration during analysis;
- ❏ *Section 5.3.3: "Changes in BTS operation"* examines how to deal with the fact that cell activity is highly dynamic, with cells are continually being added to and removed from the network. Malfunctions and consequent repairs are commonplace, and ignoring this issue can lead to incorrect analysis;
- ❏ *Section 5.3.4: "BTS service region overlap"* addresses issues which arise due to mechanisms put in place by operators to improve network performance - namely load balancing across towers with overlapping service areas. These dynamics introduce further uncertainty with regard to the recorded location and must be considered within the process of OD creation;

5.3.1 Limited precision

Challenge: The location precision of a user at the time of a network event is limited to the granularity provided by the BTS infrastructure and its geospatial layout across the extent.

Discussion & Solution: This is not a challenge per-se, but rather a fundamental characteristic of the data. Depending on the data collected by network operators, there are opportunities for geolocating handsets through available signal strength information[47,48,49,50]. However, this information is not required for billing purposes and as such very rarely recorded within CDR data. Within urban areas, however, there are typically sufficient BTSs in existence for us to derive quite fine-grained OD matrices. The density, indeed, is often greater than the granularity required for the zonal units that will underpin the OD matrix itself. Moving into rural areas, however, has the potential to result in OD matrices based on quite low-grained spatial quantization.

We recommend as **best-practice** that when interpolating from a voronoi cell representation to the representation used by the final output matrix (we use a grid based representation in Annex 1 for example), that interpolation occurs proportionately based on building counts - or even more preferably floor space coverage (due to the growing number of multi-story buildings being built in urban environments). While tower catchment areas can cover wide areas, calls predominantly come from areas where the human population congregates, and this centres round physical

[47] Most studies have relied on a data set provided by AirSage, a US based company using its proprietary Wireless Signal Extraction technology to anonymise, aggregate and analyse mobile phone signal data from multiple network operators to predict real-time traffic speeds and travel times. Similar companies include IntelliOne in the US, ITIS holding in the UK, Delcan in Canada and CellInt in Israel

[48] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination flows using mobile phone location data. Pervasive Computing, 10(4), 2011

[49] S. Jiang, G.A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M.C. Gonzalez. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In Proceedings of UrbComp'13, 2013.

[50] D. Gundlegård, C. Rydergren, N. Breyer, and B. Rajna. Travel demand estimation and network assignment based on cellular network data. Computer Communications, 95, 29-42, 2016.

infrastructures such as buildings and transport. This gives us an invaluable heuristic for pinpointing activity geospatially with greater granularity (a physical implementation is discussed in more detail in Annex 1). An example of this is illustrated in Figure 3, which shows building centroids in the port region of Dar es Salaam. Such data is invaluable in preventing network event activity being incorrectly attributed to grid-cells which predominantly contain natural features (in this case water)..

Final remark: Moreover, and as will be discussed in Section 5.5, a level of coarseness with regard to the spatio-temporal quantisation is not necessarily detrimental to analysis, and may actually be required in order to ensure privacy preserving results.

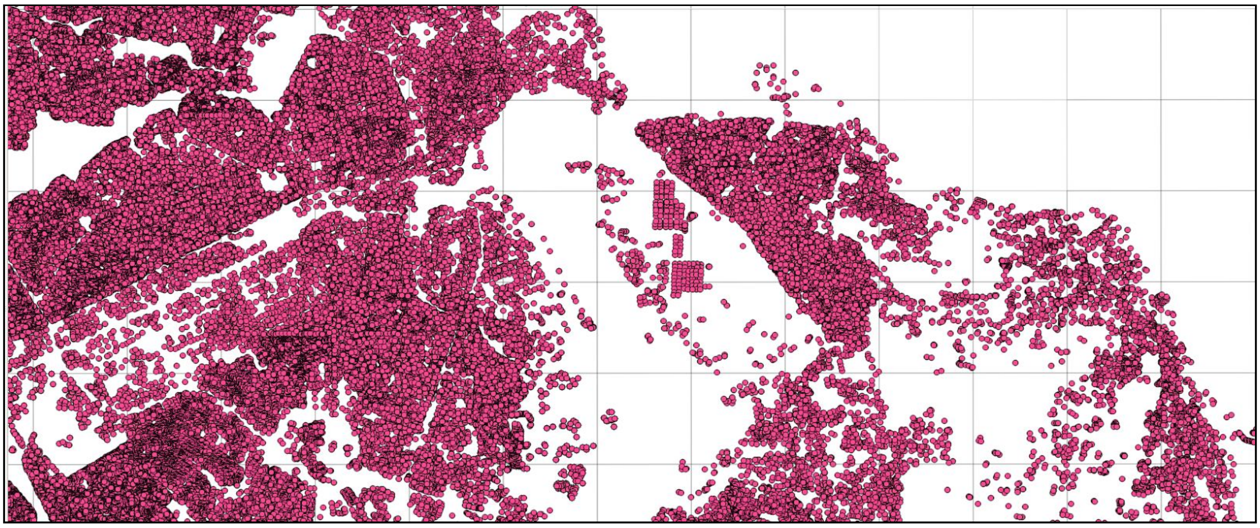


Figure 3. Example of Building Location data in Dar es Salaam Region, which provides an invaluable heuristic for performing informed interpolation of activity between voronoi regions and grid cells.

5.3.2 Non-Uniform BTS Density

Challenge: Urban spaces generally see a higher level of network activity, leading providers to operate more BTS in those areas. Due to a lower population density and associated lower levels of activity network operators therefore tend to operate fewer BTS in rural areas. This can result in BTS being located multiple kilometres apart compared to densely populated urban areas which can have a median distance of a few hundred meters between BTSs, if not less. The varying density in BTS coverage impacts the accuracy of descriptive, predictive and prescriptive insight generated outside of densely covered urban areas.

Discussion & Solution: Dealing with OD matrices with origin and destination regions of vastly varying sizes is difficult. Should one account for the size of an area when considering movement to and from it? Should there be understanding of where in a region movement is occurring from? Do variations in spatial regions affect decision making processes? Following discussions with transport experts in the UK, when constructing raw OD matrices we recommend as **best practice** the use of equally sized zonal units (which become origins or destinations respectively). Such a policy is recommended due to the ease on human interpretation that results[51].

[51] From a transport perspective within the UK zonal units are often constructed to equalise the number of trips generated per zone. Such an approach masks many of the discussed problems, with BTS density typically correlated with population and therefore journey density. The approach also implicitly addresses privacy concerns (see section 5.5). However, such an approach does introduce additional complexity in interpretation, requiring both size and value (colour) of the zonal units to be taken into account when interpreting the results (visualisation).

However, interpolation can also be misleading if not implemented with care. For example, in rural areas where there is coarse BTS granularity, and consequently voronoi cells larger than the zonal units being used, there is a danger of some proportion of journeys being attributed to unpopulated areas (This would be dangerous for example in the Mbezi ward, at the North West edge of the Dar es Salaam region illustrated in Figure 4). The computation of additional probabilistic origin/destinations on top of the derived OD matrices based on known building locations (i.e. from crowdsourced datasets[52] or from automatically identified buildings from satellite or drone imagery) is therefore particularly valuable (as detailed in section 5.3.1). In regions where there are limited areas of population, this has the potential to provide significantly better insights (although, these approaches have only been trialed within urban areas as part of this work). Understanding how CDRs can be used and relevant factors for developing rural OD matrices is currently out of scope for this work, but is a topic that provides many opportunities for future research.

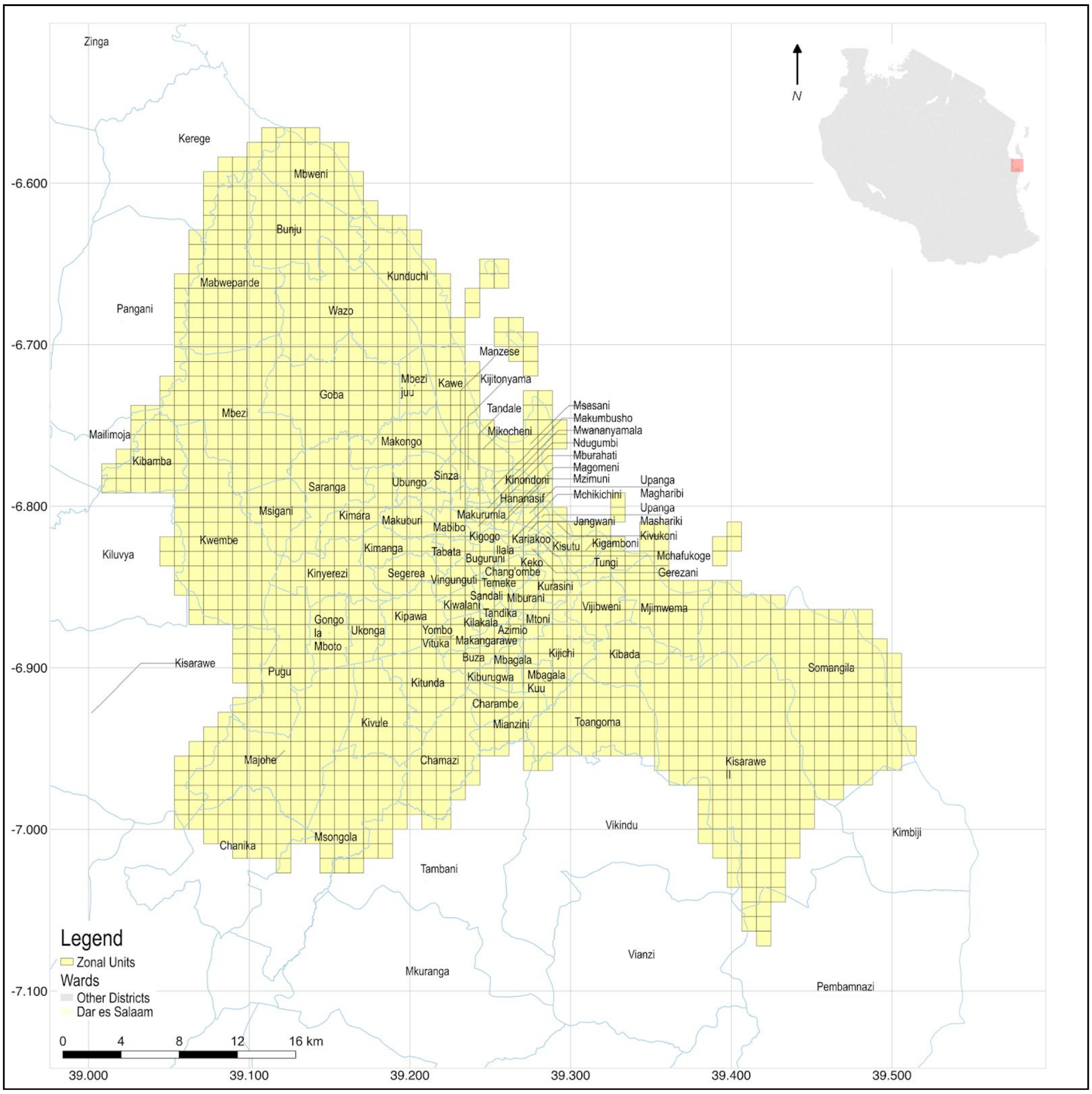


Figure 4: Example of a *grid-zonal unit representation - specifically the 1km grid used for mobility analysis of Dar es Salaam.*

[52] In Dar es Salaam such information exists due to the Ramani Huria (www.ramanihuria.org) flood resilience project

Finally, we also recommend as **best practice** the merging of BTS activity in high density areas (in Annex 1 we use a 75m cordon, below which any adjacent BTS are merged). This is due to a deleterious reduction in confidence in high density areas, which are subject to greater load balancing, that the BTS's associated voronoi region accurately reflects a user's true position[53].

5.3.3 Changes in BTS Operation

Challenge: BTS infrastructure can change significantly over time due to malfunctions and/or the installation of BTSs to meet demand. Such alterations alter voronoi cell construction (see Section 5.3) and directly affect the precision of location, with nearby towers often taking over service in the case of malfunction. Depending on how the BTS locations have been mapped to form regions within a OD matrix this can introduce significant error, with the same device location being mapped to different locations within the OD matrix. Equally the addition of BTS means that locations that were never previously reported become active at the expense of others. As such the reported locations should not be used directly as origin/destination points as they usage within the OD matrix will change systematically with the introduction and malfunction of BTS.

Discussion & Solution: In order to account for changes in BTS operation, **best practice** is to enforce that the origin/destination regions (zonal unit representation) formed as part of the OD matrix layer must be independent from the BTS locations. If only minor BTS changes are observed within the analysis timeframe then the issue can be addressed by a simple one-to-one mapping after removing BTS that do not meet an operational threshold (i.e. number of days active) and redistributing their events. Otherwise a more complex approach is required. This involves attributing journeys for a given BTS voronoi region (computed based on active BTS at the time of the journey) proportionally to all OD regions for which they overlap, based on some measure of overlap. Example measures include area, or if available, external information such as building counts. Note that the generation of such proportional mapping is often required anyway, with the generated origin/destination regions often desired to be different to the BTS regions for both commercial sensitivity, privacy (see Section 5.5) and or interpretation (i.e. visualisation by wards) reasons.

5.3.4 BTS service region overlap

Challenge: Depending on the locations of the BTS infrastructure two or more BTS service regions may overlap, within which either BTS could provide network service to a subscriber. In general subscribers will be serviced by their closest (strongest signal) BTS and therefore **best practice** is to a assume (in lieu of external data) that subscribers are located within the voronoi region of the BTS handling their network event. However, this is not always the case, with users potentially being transferred to a nearby BTS able to provide service in order to balance load on the mobile network or for other service quality reasons. As location is based on the BTS information contained in a CDR log, this effect leads to the mis-identification of a user's location, often referred to as "false movement". These rapid *handovers* occur without the knowledge of the user and can appear in CDR data in the same way as physical movement. Moreover these non-movement handovers can occur at high and varying rates across the extent, potentially generating a non-trivial amount of data that (without consideration) could be incorrectly interpreted as movement - and heavily bias OD matrices to over-represent local journeys. This is particularly true in urban areas of high BTS density.

[53] The detection of mode and purpose, and the resulting complications due to the data characteristics are discussed in Section 5.4

Discussion & Solution: An effective way to address issues of false movement is within the stop identification stage of the OD matrix creation. **Best practice** recommendation is to address this issue by preventing artificial stops being identified when such events are occurring. As detailed in section 5.2.1, this is achieved by the requirement for a stop to consist of a minimum two or more consecutive co-located events separated by a minimum (parameterised) time period. This is also part of the mechanism utilized to address sparse temporal frequency bias, sharing the same caveats and justification. Most previous studies have employed a minimum temporal filter of 10 minutes between records to account for false displacement[54,55], however determining this threshold from ground-truthing is the optimal approach if possible. In addition to this, we also recommend application of a low *minimum inter-stop time* (set to <2 minutes in the implementation in Annex 1), which directly removes handovers from being considered journeys (n.b. if this minimum inter-stop threshold is set to zero, we will again return to a transient rather than a stop based OD matrix.

An alternative solution is the exclusion/inclusion of surrounding BTSs as part of the stop detection process (as these are the towers where load balancing will occur). Due to the non-uniform BTS density (see Section 5.3.2) such an approach is only applicable in densely served environments, however, and has generally only been attempted with geolocated CDR data with a spatial filter between 300m and 1km based on accuracy of device collecting location data[56]. Specifically an exclusion approach would simply ignore events from surrounding towers when looking for a subsequent co-located consecutive event to fulfil the definition of a stop, while an inclusion approach would count surrounding BTSs as being co-located[57,58,59]. The impact of this approach, however, currently requires further investigation.

5.4 Lack of directly recorded movement metadata

Challenge: Traditional OD matrices constructed via Roadside Interviews (RSI) or other survey based approaches typically contain additional meta-information such as trip purpose, vehicle type/transport modality and vehicle occupancy. This data is not explicitly available from CDR data.

Discussion & Solution: Despite not being directly available, it remains possible that in the future algorithmic improvements will allow meta-information to be inferred from journey data via the integration of assumptions and/or additional data sources. The best methods to do this are necessarily specific to the metadata required - so no specific best practice can be provided for all potential metadata. However, even when inference is employed, a general **best practice** is to perform that analysis in a supervised fashion and not solely rely on heuristics (such as movement speed). This means obtaining a sample of ground truth metadata via roadside interviews and/or

[54] F. Calabrese, G.D. Lorenzo, L. Liu and C. Ratti. Estimating Origin-Destination Flows Using Mobile Phone Location Data. Pervasive Computing, 10(4), 2011.

[55] Md. S. Iqbal, C.F. Choudhury, P. Wang, and M.C. Gonzalez. Development of origin-destination matrices using mobile phone call data. Transportation Research Part C, 40, 2014.

[56] Most studies have relied on a data set provided by AirSage, a US based company using its proprietary Wireless Signal Extraction technology to anonymise, aggregate and analyse mobile phone signal data from multiple network operators to predict real-time traffic speeds and travel times. Similar companies include IntelliOne in the US, ITIS holding in the UK, Delcan in Canada and CellInt in Israel

[57] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination flows using mobile phone location data. Pervasive Computing, 10(4), 2011

[58] S. Jiang, G.A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M.C. Gonzalez. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In Proceedings of UrbComp'13, 2013.

[59] D. Gundlegård, C. Rydergren, N. Breyer, and B. Rajna. Travel demand estimation and network assignment based on cellular network data. Computer Communications, 95, 29-42, 2016.

camera-based analysis at key locations from across the region being analysed (with the number of interviews and locations deployed on a best effort basis to achieve a feasible accuracy cost trade-off). Nonetheless this approach remains an open machine learning research task.

In order to provide *trip purpose* metadata, **home** and **work** locations must first be inferred (and indeed, this can be achieved in a relatively robust fashion). This enables the subsequent identification of journeys between home, work and other locations[60]. Inference of these location types can be done jointly based on measures of occupancy frequency and consistency (BTS usage entropy) during hours known to correlate with times people are at home and work[61,62]. While such inferences can never be 100% accurate (i.e. night workers will be incorrectly inversely labeled), this approach provides insights of high utility at limited cost when performed at scale.. Note that further refinements could be undertaken (i.e. to identify night workers based on known areas of night work), however, this attempts to return the definition of home and work to a more informal vaguely defined concept which may undermine explicit insights, and this we recommend maintaining a focus on day and night BTS modes for each anonymised individual.

Provision of transport *modality*, in contrast, requires the CDR data to be augmented with additional information. Here inference is complex, and requires extensive further research when performing on CDR data. Hence we recommend a **best practice** of targeted sampling of locations for roadside interviews to determine the proportion of different modalities. This can then be linked to estimated traffic counts over the same road/path segment based on a given OD matrix by utilizing routing algorithms over external transportation network data[63]. These, in turn, then provide a scaling factor at each location from which to arrive at population values. Error can hence be measured and mitigated with increasing confidence in the number of locations (repeatedly) sampled on a best effort basis trading-off confidence and cost. Again in the future the requirement for roadside interviews may be reduced, and instead replaced with computer vision algorithms utilizing fixed roadside cameras and/or satellite or drone imagery.

5.5 Privacy and Ethical Implications of Using Call Detail Records

Challenge: CDR data contains significant information regarding an individual's behaviour. The creation of OD matrices from such data is therefore required to be done in such a way to as not violate individual's right to privacy including, but not limited to, meeting any legal or regulatory obligations.

Discussion & Solution: In emerging economies limited ethical, legislative and regulatory frameworks exist, if at all. Certainly, **best practice** recommendation is to use the guidelines[64] from the GSMA as a

[60] Further location types have been considered as part of activity-based approaches but found to be causing ambiguous relationships between mixed-land use and activity types. Trip-based (i.e. stop or transient) are generally chosen over activity-based approaches due to the aforementioned relationship and complexity of implementation involved. Further reading for activity-based approaches can be found in Transportation Research Board. Activity-based travel demand models: A primer. Strategic Highway Research Program 2015.

[61] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data, PLoS ONE, 7(6), 2012.

[62] M. Dash, H.L. Nguyen, C. Hong, G.E. Yap, M.N. Nguyen, X. Li, S.P. Krishnaswamy, J. Decraene, S. Antonatos, Y. Wang, D.T. Anh, and A. Shi-Nash. Home and workplace prediction for urban planning using mobile network data. Mobile Data Management (MDM), 2014 IEEE 15th International Conference on2, pp. 37–42, 2014.

[63] For instance from openstreetmap in Dar es Salaam

[64] Developed following the Ebola epidemic in Western Africa, the guidelines are specifically for the usage of CDR data in *"exceptional circumstances"*. However, it is the experience of the DECs team that these guidelines are the de facto guidelines used regardless of situation. The guidlines are avaliable from:

fallback to this situation (most network operators are GSMA members). Of relevance to the process of OD matrix creation are the guidelines that require *only* network operators have access to non-anonymised CDR data: that analysis will not be undertaken that could single out identifiable individuals and that the resultant output is non-sensitive.

The first of these guidelines is met without compromise, as OD matrix creation does not require the identity of any individual at any time, only aggregated movement patterns. One concern may be that in some cases the sparsity of the data may enable the movement patterns to be re-attributed to a specific individual using external information. This concern is ameliorated by both the granularity of the data, location granularity is broadly and implicitly limited to an area covering significantly more than one person due to BTS placement, and the vetted nature and agreement of the individual's undertaking the analysis to not attempt such re-attribution. This latter requirement on processing additional satisfies the second requirement.

The final requirement, that the resultant output is non-sensitive requires additional checks on all produced OD matrices and their associated metadata. Here, our best practice recommendation is use of k-anonymity, a common method to producing non-sensitive output. Specifically, *k*-anonymity, ensures that any individual cannot be distinguished from at least *k-1* other individuals[65]. Since OD matrices do not store information about individuals, but rather counts of journeys between if you know the time someone left a zonal unit, to preserve k-anonymity you need to ensure k-1 other people left the area over the period being considered to preserve privacy[66]. Now, as these technologies develop, presented periods will likely increase in granularity, with k-anonymity placing constraints on what can be exposed to ensure confidence that privacy is being maintained. This is of course always an arbitrary choice - others would no doubt be content with a much lower k value, however given the exploratory nature of this work and general novelty we advise to err on the side of caution.

regions this involves ensuring that counts within each cell are greater than *k x m,* where *m* denotes the maximum number of journeys for any given individual. I.e. that certain journeys could not be attributed back to any single *k-1* people as the only people undertaking such journeys. This represents an easily computable upper-bound to preserving privacy in this fashion. However, for large temporal periods this may result in significant amount of data being omitted. An alternative, when available and a significantly large temporal period is reported, is to utilize known population statistics of the regions[67], ensuring that the population in each region is greater than *k*. While users can potentially be identified through their preferred location visits, this risk is significantly reduced when using CDR data due to the coarse spatial granularity involved[68]. Traditional approaches to developing OD matrices in the UK typically enforce 12-15 journeys per region to ensure privacy, suggesting a similar value for *k* would be appropriate.

http://www.gsma.com/mobilefordevelopment/country/global/gsma-guidelines-on-the-protection-of-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-ebola-outbreak

[65] B. Gedik, and L. Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Transactions on Mobile Computing, 7(1):, 1-18, 2008

[66] Note, in the outputs of the accompanying mobility report a whole year is used as the aggregate period, so the minimum granularity we have to address is "weekends" (representing 104 days out of 365 days in the year). The amount of journeys for even the zonal unit with the lowest movement is ~1000 vehicular journeys, providing a k-anonymity of ~1001, giving extensive confidence in privacy preservation.

[67] In the implementation part of this work this is approximated by building counts as no reliable population estimates were available. In this case *k* represents a requirement that journeys can not be reattributed to individuals frequenting/residing in *k-1* buildings.

[68] Y. de Montjoye, C.A. Hidalgo, M. Verleysen, and V.D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, vol. 3, 2013.

6. An Agenda for Future Research

Through identifying the state of the art in OD creation via CDR data, further avenues of research have been identified that will push mobility analysis into its next stages of effectiveness, especially in emerging economies. These are listed below (although are not laid out in any order of priority).

Promising Research and Development areas

6.1 Country/Region-wide Origin Destination Matrices

As demonstrated in this work, research into using CDR for OD and mobility analysis primarily occurs in cities. However, it has become clear in this analysis that it would only take limited research to expand these processes to produce mobility maps for **all of the large urban areas of Tanzania** - and indeed the whole country, including **rural areas.** Additionally, it was observed that a highly significant number of journeys passing through Morogoro represented destinations at industrial zones within the city, indicating analysis of country-wide **industrial transport flows** may also become possible.

6.2 Dynamic Generation of Demographic Maps

National statistics Institutions across the developing world are often poorly resourced, with insufficient capacity to meet the demands of rapid urbanisation and population growth. The effect of this for cities such as Dar es Salaam, means that demographic information is sparse if available, though commonly is non-existent or not temporally accurate. Even when it is collected it is likely to be rapidly out of date. It is our strong belief that **CDR data can reveal these demographics** (when combined with state of the art machine learning techniques, but which are readily available), extending OD mobility maps to **full demographic information maps**, that are data driven and hence updateable at low logistical cost every quarter. Examples of the new (neo) forms of demographic informations that this could add to the advanced mobility insights gleaned here, are insights into: **household expenditure** (through mobile money); estimates of **health and educational levels**; **employment estimates**; dynamic analysis of **unplanned urban expansion**; predictions of **population growth.**

6.3 Advanced Mobility Analysis Via Computer Vision/Drone Technology

Ground truth measurements sampled across the road network are essential for correctly scaling the generated OD matrices and correcting for bias. Collected by traditional techniques these are time consuming and only a minimal number are taken. However, such counts can be done semi-automatically through the huge advances recently made in **machine vision** (and "deep learning"). Employment of **cameras/drones** in combination with these computation techniques will allow *significantly* more samples to be taken on an ongoing basis. This increase in sample size and temporal granularity would enables the generation of much higher confidence levels in resulting matrices - and provides the opportunity for low cost **continually updated** matrices. Such analysis would also then significantly improve **transport mode analysis** when combined with CDR data, as well as supporting demographic analyses detailed in 6.2, as well as further infrastructural issues (such as road condition analysis).

6.4 Advanced Mobility Insights via Topic Modelling

While CDR data allows for extensive developments in the generation of OD matrices, the extra fidelity it brings can actually complicates issues of interpretation and visualization (for example in Dar, prior analyses have concentrated on around 10 super-regions and flows between them, which made sense given the precision and size of the datasets being used). Topic Modelling is a relatively new technique used to summarize complex datasets for purposes of analysis. It is based on a mathematical approach called Latent Dirichlet Allocation that, if applied to OD matrices, would **reveal underlying movement trends** and **key transport patterns** occurring across the extent that were: 1. Previously lost inside the data, even when contained in OD matrices; 2. Directly translatable into transport policy insights; 3. Able to provide new visualizable descriptions of the the city.

6.5 Financial Flow Analysis

CDRs can include data on Mobile Financial Services, with associated location. Accordingly, the mobility of money across a city can now be identified. From this point, it would be possible to identify not just the 'home' areas of the population, but also where money is spent - and vitally how that **money flows** around the city, the region, and whole country.

ANNEX 1: Physical Implementation of Origin-Destination Matrix Generation from CDR data

Having introduced the nature of CDR data, and the the opportunities and challenges that emerge in using it as a cornerstone for mobility analysis, we now present a practical and technical description of how we implemented the approach (and specifically addressed those challenges for the Dar es Salaam case study region). The technical process of how OD Matrices were generated from CDRs data is described in 7 steps below (each step corresponding to the high level summaries that were listed in section 4.2):

1. **Data Preparation:**

   a. **Raw Data Collation:** The first stage of analysis required that 5 network event tables were extracted from the raw CDR data tables shared for purposes of mobility analysis by the network provider. A high performance SQL instance is sufficient for these purposes:
      → incoming call events;
      → incoming sms events;
      → outgoing call events;
      → outgoing sms events;
      → data events;
   As the region under consideration was Dar es Salaam, each table either listed all events where the initiating tower was located in the Dar Region (outgoing events) or the receiving tower was in Dar es Salaam (incoming events). Note that this means individual communications can appear twice, but with a different anonymized subscriber id in each case (the initiator and recipient respectively). Any subscriber who does not visit Dar es Salaam at least once is filtered out.

   b. **Master Event Table:** An SQL union of these tables (UNION ALL) was then taken to produce master network event table, keyed by the anonymized subscribed id and a timestamp.

   c. **Distinct Subscribers:** A list of distinct subscriber ID existing in this combined table could then be extracted - this list represents all anonymous subscriber ids observed as being active in Dar es Salaam at least once over the period (totalling just over half a million individuals in our selected sample).

   d. **Data Cleansing:** At this point both any BTS and spatial regions were merged as follows:

      i. **Proximity Merging:** In order to deal with the intermittent BTS any towers not active for at least than *300 days* (82% of the year) had their activity with the nearest fully active tower, based on euclidian distance (this is a direct implementation of the solution detailed in section 5.3.2);

      ii. **Proximity Merging:** In order to deal with sets of BTS located in such close proximity that the location of user cannot be distinguished the activity of all towers within *75m* of each other was merged with that of the most active tower in the set. (this is a direct implementation of the solution detailed in section 5.3.3);

   e. **Event Series Extraction:** Just over half a million event series were then produced by filtering the master event table for each of the distinct anonymized subscribers occurring.

2. **Stop Sequence Generation**

   a. **False Movement reduction:** At this point any events occurring at a new tower that occur within a duration, ***s***, of a prior event at a different tower are assumed to be load balancing artifacts and removed from the dataset. This was set at a very conservative value of *2* minutes to ensure protection of false movement (a direct implementation of the solution discussed in section 5.3);

b. **Stop Identification:** A distributed python script was then used to convert each event series into a sequence of valid 'stops'. A stop is a set of at least **k** contiguous events which are all recorded at the same BTS over a period longer than duration, **d.** The maximum gap, **g**, between events (known as the *max_inter_event_time)* is also specified (n.b. hours between 1am and 6am are not included in this gap time as they reflect sleeping periods where minimal network activity occurs). The parameterizations used in the Dar es Salaam case study are listed in Table 3:

Item	Symbol	Value	Description
minimum number of events	k	2	The minimum number of consecutive events required within a given region for the segment to be considered a stop.
minimum permissible duration	d	10 mins	The minimum permissible duration between the maximal and minimal times of a set of events with the same tower_id to consider them a stop.
maximum inter-event gap	g	4 hours	The maximum time between any two network events before they are considered to be non-consecutive events due to the high probability of unobserved movement.
minimum inter-event gap	s	2 mins	The minimum time permitted between events. This is set to prevent detection of false movement from identified stops, due to handovers or other network load balancing.

Table 3: Parameterization of the stop identification process for mobility analysis of the Dar es Salaam region.

c. **Confidence Assessment:** All stops were committed to an SQL table along with a value representing our confidence in each stop's efficacy. This confidence score was calculated as follows one minus the ratio of the largest event gap to the whole duration period of the stop.

3. **Journey Generation**

a. **Journey Identification:** Stop sequences were then converted into a set of journeys for each subscriber. A journey is defined as 2 contiguous 'stops' that are separated by a period of at least t_{min}, but no more than t_{max}. The value for t_{min} ensures that false movement hasn't slipped through the net(a double check of the solution discussed in section 5.3). Setting t_{max} ensures that we are unlikely to have reconstructed a journey that, in actually, is missing a midpoint destination.

b. **Confidence Thresholding**: Journey's which do not meet a predetermined confidence threshold may also be removed at this point. Confidence is determined as the mean of the respective confidences attributed to the journey's origin stop and destination stop - and in this instance was enforced to be greater than 0.1. Note this favours journey's whose origins and destinations contained more events occurring over a longer period.

4. **Journey Cleansing**

a. **Outlier Removal:** The dataset is then filtered to remove individuals with outlying behaviour (i.e. disproportionately high network usage) who would hence skew the representative nature of the results due to the disproportionately large number of journeys attributed to them (this is a direct implementation of the solution discussed in section 5.2.2). The threshold for outlying behaviour, j_{max}, is subjective. Following observation of the distribution of journey frequencies over the year, a value of 4000 was selected. We hence filtered out all subscriber who were detected as having undertaken more than 4000 journeys over the year (over 10 per day - one would expect many of these users would be driving public transport).

5. <u>Metadata Tagging</u>

   a. **Day/Night mode detection:** In order to detect journey purpose metadata, and allow us to categorize journeys as either: **Home Based Work (HBW); Work Based Home (WBH); Home Based Other (HBO): or Non Home Based (NBO);** the CDR data was processed to identify subscribers where we could, with high confidence, designate their Home and work based BTS. This was achieved by identifying the tower that they used the most at night and in the day respectively (night/day modes).

   b. **Filtering:** Once modes were obtained for each subscriber, those who did not meet a sufficient threshold of observations, or whose normalized entropy[69] (essentially their variance in tower usage) was too high were discarded. his was parameterized as detailed in Table 4 below:

Item	Symbol	Value	Description
min observations	o	50	The minimum number of observations that the we must have had of the individual at the designated mode BTS.
maximum normalized entropy	H	0.5	A representation of the spread of towers used by a subscriber over the specified period. A high entropy means the individual is seen at a higher variation of towers, which in this instance is undesirable as we are less confident as to *which* is the person's true home or work location.

Table 4: Parameterization of the day and night mode identification process for journey purpose tagging

   c. **Tagging:** Finally all origins and destinations for just under 200,000 subscribers in our sample who survived this process were tagged with a home or work cell as appropriate, hence allowing categorization of a subset of journeys into HBW, WBH, HBO and NBO types.

   d. **Aggregation:** This process additionally allowed us to visualize the distribution of residences and workplaces across the extent. This is illustrated in Figure 5, however far more detailed renderings and population statistics are supplied in the accompanying report "Mobility Insights in Dar es Salaam".

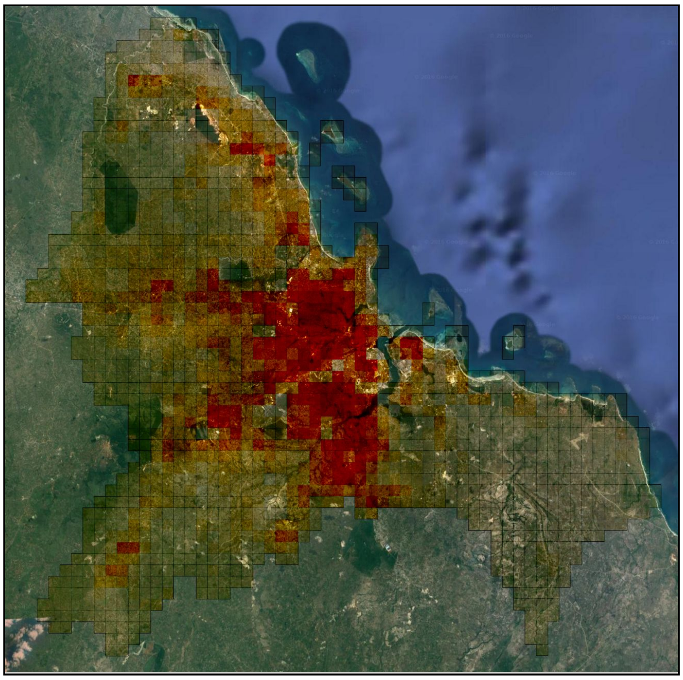


Figure 7. Visualization of the number of homes detected in each cell, across the Dar es Salaam case study extent.

[69] normalized entropy provides an indication that the subscriber didn't settle in one place for the majority of their time and therefore we are unable to predict that the most common location was their home/work cell with sufficient confidence.

6. **Intermediary OD Generation**

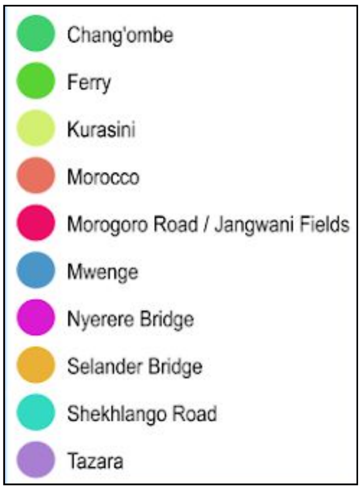
   a. **Journey Aggregation:** Finally all anonymized subscriber information was completely removed from the entire data, and a union taken of all sets of journeys across the subscriber base. This produced our entire journey set for the period examined.

   b. **Matrix Generation:** From this journey set a corresponding OD matrix was created (via a simple process of counting the number of journeys identified between each BTS). The input journey set used was also varied through filtering to allow various forms of analysis: time of day, day of week, time of year, minimum journey duration, minimum journey distance or journey purpose (based on work, home or other), etc. However, at this point journey counts are restricted to our sample size.

7. **Scaling**

   In order to convert our intermediary, relative OD matrices into absolute counts representing the whole population, a process of scaling has to be implemented. This process as implemented in the Dar es Salaam case study is described below:

   a. **Obtaining Ground Truth Traffic counts:** Raw OD Matrices must then be scaled in order to correctly extrapolate the number of detected journeys to the full population. By necessity, this process requires some external ground-truthing data. This was performed using video camera based traffic counting at the multiple traffic sample points listed below (please see Figure 1 for a map of these points).



● Chang'ombe
● Ferry
● Kurasini
● Morocco
● Morogoro Road / Jangwani Fields
● Mwenge
● Nyerere Bridge
● Selander Bridge
● Shekhlango Road
● Tazara

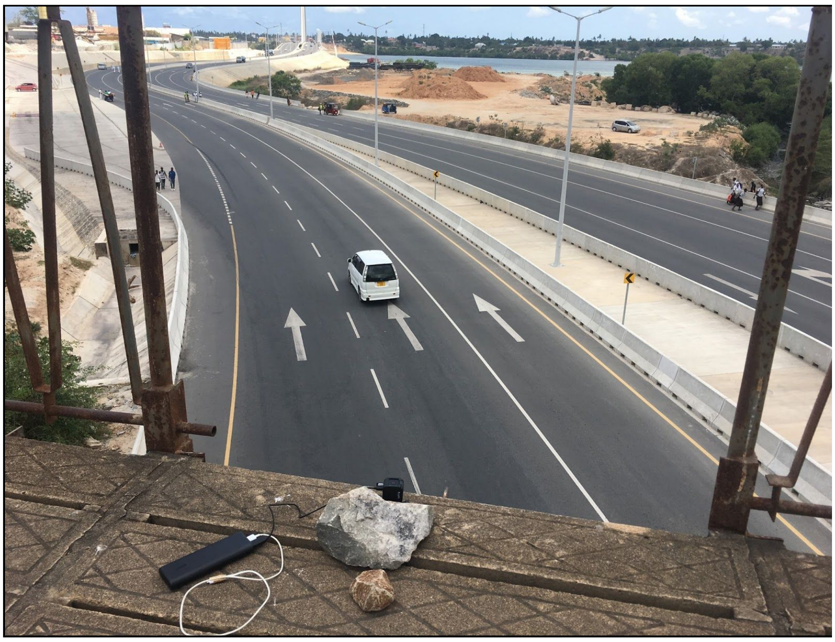


Figure 8. Chosen survey points

Figure 9. Example setup of traffic camera

The points of Figure 8 were chosen due to their key locations on arterial roads in consultation with with the Dar es Salaam Master Plan team with attendees of Dar es Salaam City Council and JICA. The collection of videos was conducted in a manner similar to that identified in Figure 9.

   b. **Calculating Scaling Factors:** In order to determine a scaling factor ground truth traffic counts were taken at ten different locations across Dar es Salaam (incoming and outgoing traffic for 8 locations, incoming only for 2 with an average of 6 hours over three periods in the day, for multiple days). All locations selected were major roads. For each location a scaling factor was computed as follows:

i. Computing the route between every region pair using the OpenStreetMap road network information modified to include the Kigamboni ferry. For each region routing was undertaken from the closest routing point to the center of the region. If the route taken included the observed road segment then the journeys between the regions as indicated by the OD matrix were attributed to the estimated journey count. In order to get a robust estimate of the scaling factor the average journey count for a day (24 hrs) over the year was taken;

ii. The observed data was extrapolated to an estimated count over the 24 hour period and scaled up to represent all journeys (in contrast to just motorised transport) using a factor based on the proportion of motorised to non-motorised transport reported in the Road Side Survey of 2007 conducted by JICA., in combination with their reported statistics for passenger ratios for different modes of transport.

iii. A scaling factor was then computed as the multiplier required to adjust the OD estimated observation count at each location to the figure based on the observed data. A global scaling factor was then computed as the mean of these;

8. <u>**Output OD Generation**</u>

a. **Determining a Zonal Unit Representation:** Finally, results could be interpolated into the specific output geospatial representation. This serves to protect both individual and commercial privacy, with matrices being projected to a disjoint set of geographical regions (zonal units) that underpin the final output OD matrix and match the task at hand. To move from voronoi cells to grid 'areas' first a grid of dar es Salaam was settled upon. This grid contained cells of 1km^2, covering a wide spatial extent and broadly covering areas where buildings or transport infrastructure evidence human population activity. Any grid-cells that were generally determined "building less" were deemed safe to be removed from analysis and/or merged into external regions.

b. **Fine Tuning the Representation:** The grid was clipped to the shape file of the country (providing a smooth coastal outline), and clipped to only incorporate wards with the Dar es Salaam administrative region. A selection of 30 external zonal units were then also manually crafted to allow a non-grid addition to the zonal unit representation for: 1. Main through routes in the Dar es Salaam region; and 2. To reflect representative BTS across the rest of the country. Much of this process occurred within the QGIS environment, resulting in production of the regional-representation.

c. **Interpolation:** conversion of journey counts from the voronoi cell representation to this new hybrid 1km^2 grid + external zonal unit representation occurred as follows. First the grid was "sharded", layering the grid and voronoi representations on top of each other and 'cutting out' areas that intersect (i.e. the shards). This produces a grid/voronoi cell 'cover' mosaic, as illustrated in Figure 11 (where darker colours reflect the amount each component of a grid cell is representative of it's parent voronoi cell. The darker the component is the more it is 'sucking up' the BTS's data).

A weight can then be attributed to each shard, reflecting how much of its parent voronoi cell it represents. This weighting could be based on a number of methods:

- the proportion of spatial intersection;
- the proportion of the voronoi cell's overall building counts the shard intersects with;
- the proportion of the voronoi cell's overall floor space the shard intersects with;

In our analysis we use building space to calculate a shard's weighting factors (as per the solution recommended in Section 5.3.1).

For scalar statistics (such as number of residences estimated in a cell), a figure for each grid-cell can then be calculated by 1. first estimating a value of the statistic by multiplying the count for the shard's parent voronoi cell by its weight; and 2. adding up the resulting values for all shards contained in the grid cell. However, for journey interpolation, things are slightly more complicated. A value for the number of journeys between each <origin-cell, destination-cell> pair can be calculated by:

i. Estimating all <origin-shard, destination-shard> counts, by multiplying the journey count *between* both shards parents by *both* shards weightings;

ii. Summing these counts for every shard contained in the origin and destination grid cells;

Once calculated any cell to cell route can be constructed in exactly the same manner, working out the shard to shard weighted (multiplied) journey counts it contains, and adding them together. This produces a new OD matrix for each voronoi, in grid-cell format, as illustrated in Figure 12.

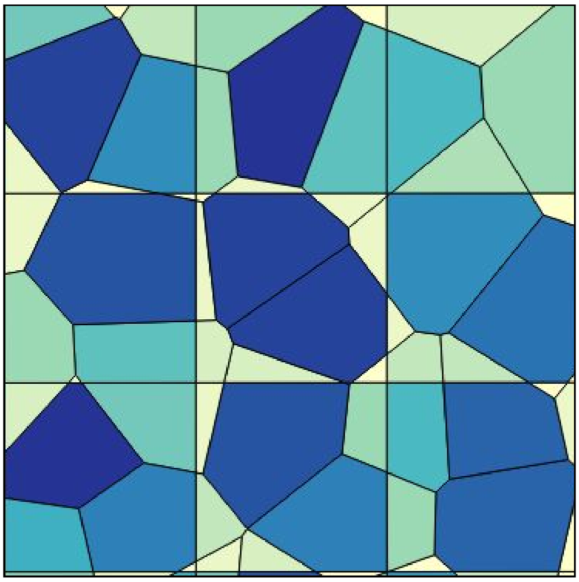


Figure 10: A synthetic example of Sharding to allow interpolation between voronoi cells and grid cells. Here the shade of the shard reflects the percentage of buildings the shard covers relative to the whole voronoi cell the shard intersects with. This provides a weighting that can then be used to interpolate with.

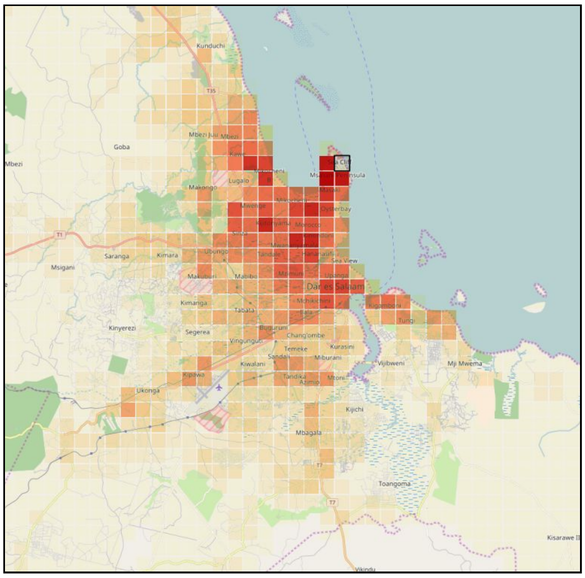


Figure 11: An example of an OD matrix once it has been interpolated into the grid-based zonal unit format (via the shard format intermediary illustrated in Figure 6). This OD map represents destinations of journeys leaving the masaki peninsula area of Dar es Salaam.

d. **Collation:** by applying this interpolation process we produced 9 different OD maps for each of the 1527 zonal grid-units (a total of ~14,000 maps). Each analysis visualizes 9 different OD maps for the focus zonal unit, incorporating representations of:

- inbound journeys;
- outbound journeys;
- commuting patterns to and from the zonal unit:
- temporal breakdowns;
- summaries of inbound traffic from external regions;
- impact on transport flows via routing densities.

A much richer description of these results is providing in the accompanying report “Mobility Insights in Dar es Salaam”.

ANNEX 2: Additional Map Generation from CDR data

Following the methods of core OD Matrix generation described in Annex 1. Numerous insights and maps can be subsequently created from the data.

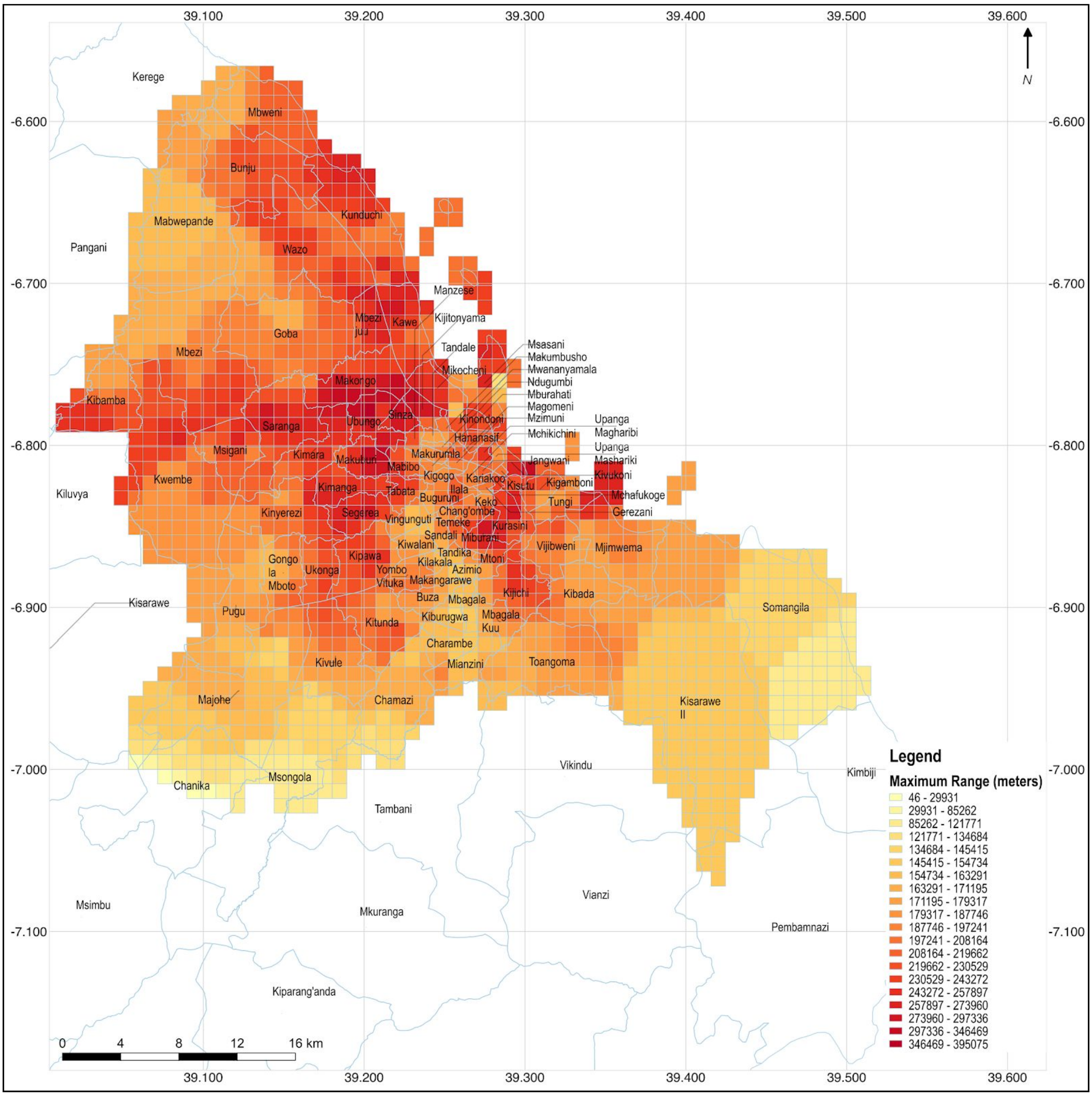


Figure 10: Shows the Mobility Across Dar es Salaam.

Mobility Map Generation

The median and maximum range of individuals whose homes were identified within each BTS's catchment area. These are then aggregated at the BTS level and converted into a grid-cell representation to produce two different forms of mobility maps.

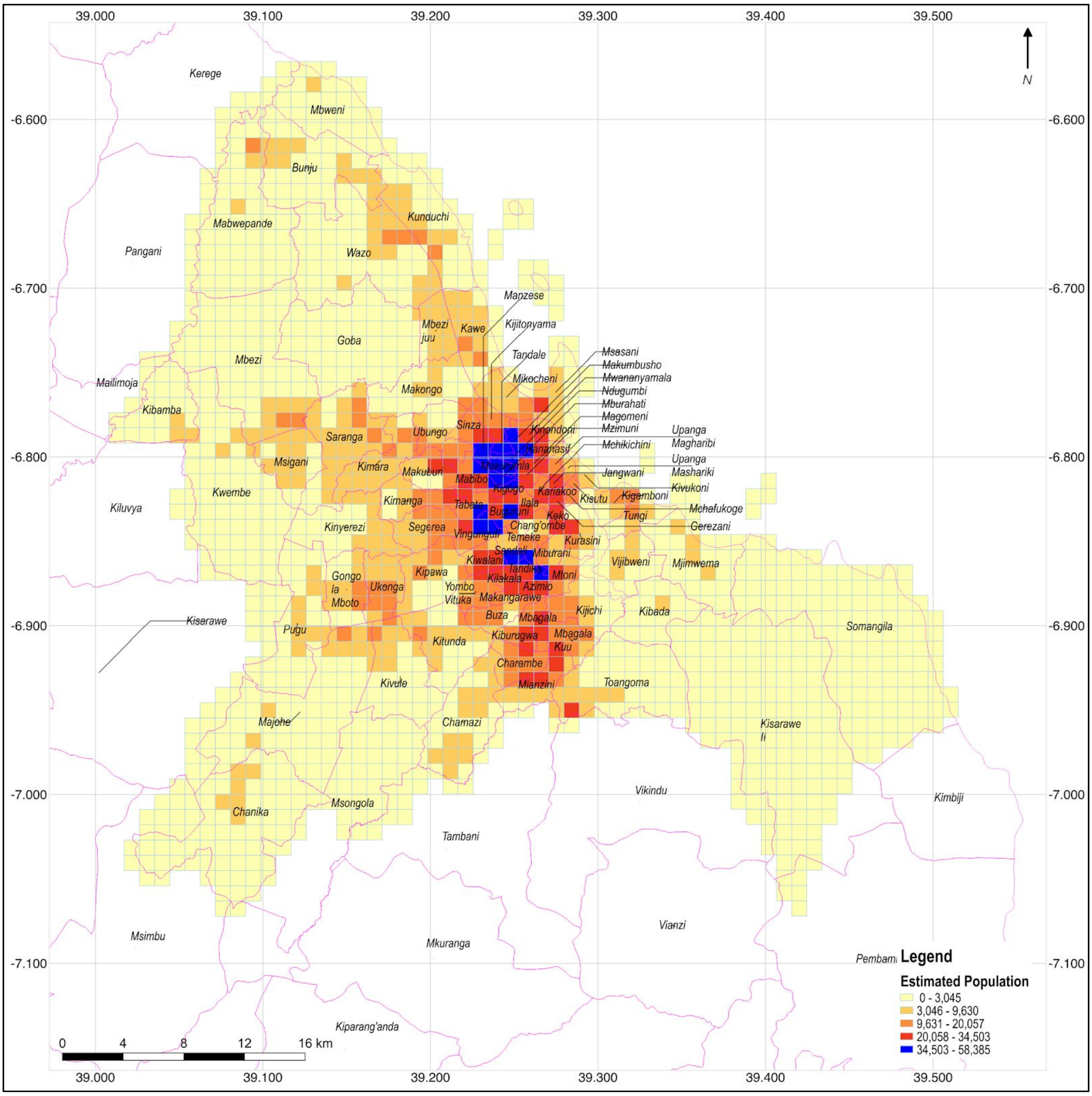


Figure 11: Population statistics derived from a CDR information

Population and Activity Map Generation

SMS and Call events give an indication of the population of an area. However, so do the number of homes attributed to each tower. Counts for each of these can be aggregated at the BTS level and converted into a grid-cell form to form activity and population maps respectively.

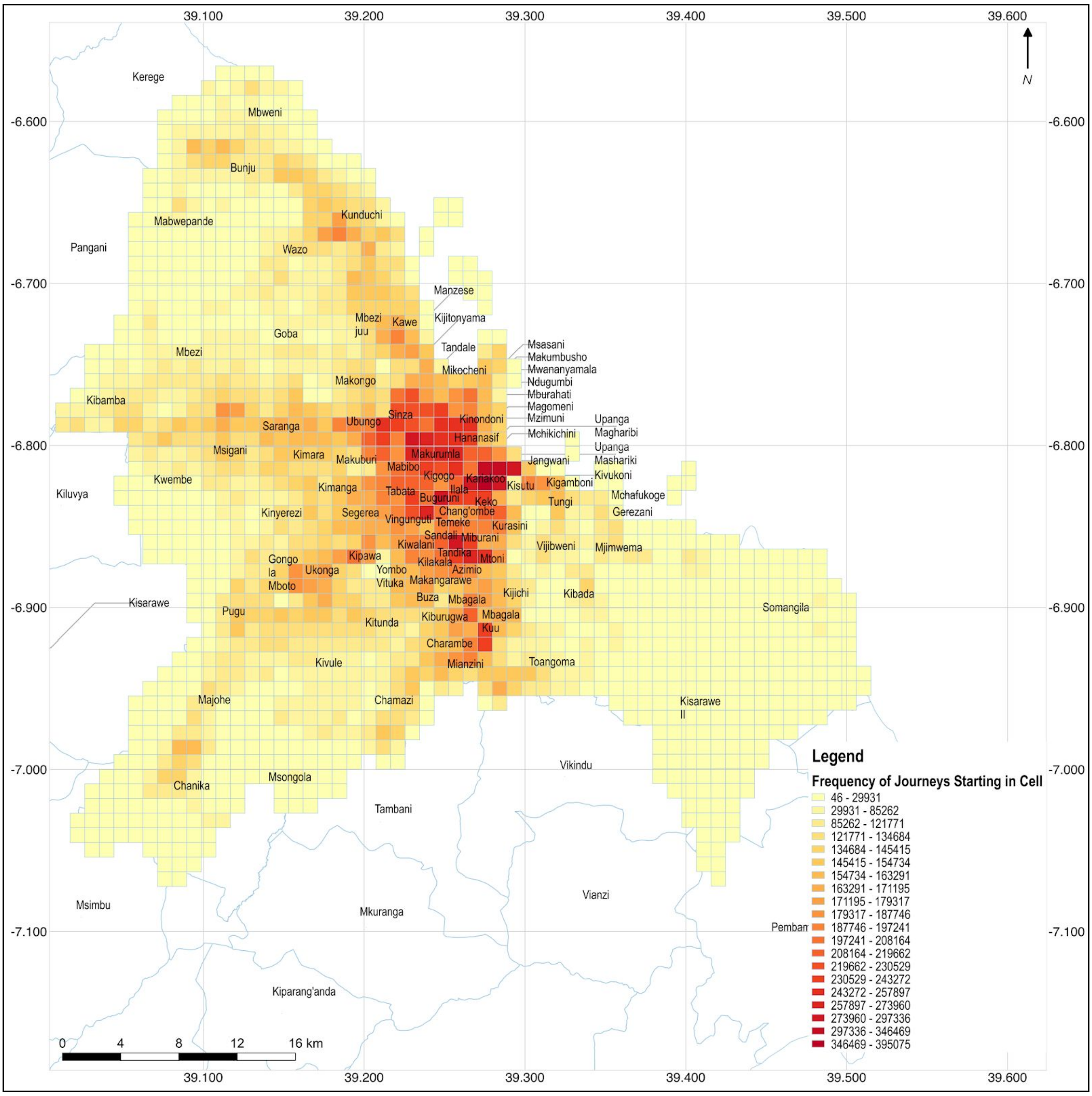


Figure 12: The Frequency and Flow of Journeys for Each Zonal Unit

Through Flow Map Generation

By examining the number of different individuals who appear in each zonal unit, or even better the entropy of each zonal unit in terms of those individuals (hence detecting transience), you can distinguish between which areas are static (in terms of the people who are active there) and which have high throughput to produce variations of through-flow maps.