Measurement of Human Mobility Using Cell Phone Data: Developing Big Data for Demographic Science*

Nathalie E. Williams[1], Timothy Thomas[2], Matt Dunbar[3], Nathan Eagle[4] and Adrian Dobra[5]

Working Paper no. 137
Center for Statistics and the Social Sciences
University of Washington

November 19, 2013

[1] Department of Sociology and Jackson School of International Studies, University of Washington
[2] Department of Sociology, University of Washington
[3] Center for Studies in Demography and Ecology, University of Washington
[4] School of Public Health, Harvard University
[5] Department of Statistics, Department of Biobehavioral Nursing and Health Systems, Center for Statistics and the Social Sciences, and Center for Studies in Demography and Ecology, University of Washington

* This paper will be presented at the 2014 Population Association of America Annual Meeting, Boston, MA, May 1-3, 2014.

Abstract
Human mobility, or movement over short or long spaces for short or long periods of time, is an important yet under-studied phenomenon in the social and demographic sciences. The rarity of research on mobility is likely due to the difficulty in measuring mobility with surveys and censuses. In this paper, we use a different data source, cell phone call records and develop, test, and compare five measures of mobility. Cell phone data provide a unique and unprecedented opportunity to study mobility around the world and with this data source, we are able to measure mobility, or daily movements of short or long duration and short or long distances, of around 1.5 million people living in Rwanda between 2005-2009.

Introduction

Human mobility, or movement over short or long spaces for short or long periods of time, is an important yet under-studied phenomenon in the social and demographic sciences. With possible associations to social interaction and networks, the spread of infectious diseases, economic well-being, growth, and change, and migration, mobility is likely a key factor to many social behaviors and macro-level change. In addition, along with health and marriage, mobility stands to contribute strongly to the key sources of demographic change: fertility, mortality, and migration. While the potential contributions of mobility to understanding social and demographic behaviors are immense, the study of this particular behavior remains rare. This is likely due to the difficulty in measuring mobility with surveys and censuses, which are the primary tools of quantitative social scientists. Thus, development of new ways of measuring mobility should be a high priority and could lead to significant advances in the social and demographic sciences.

In this paper, we develop, test, and compare five measures of mobility derived from cell phone call record data. Cell phone data provide a unique and unprecedented opportunity to study mobility around the world, with an estimated 128.2 and 89.4 cell phones per 100 people in developed and developing countries respectively (Citation). With this data source, we are able to measure mobility, or daily movements of short or long duration and short or long distances, of around 1.5 million people living in Rwanda between 2005-2009.

In order to further interrogate our mobility measures, we apply them in Bayesian spatio-statistical models that allow us to assess their efficacy in understanding relationships between contextual factors and mobility. These models combine spatially and temporally detailed contextual data such as population density, and accessibility to roads, the Rwandan capital Kigali, and international borders. These tests reveal two key results; first, that our mobility measures perform well in the modeling context for which they are ultimately intended, and second, that our mobility measures perform better or more accurately when combined with contextual data.

Background- Current Status of cell phone data and mobility measurement

Note to PAA reviewer:

This section will contain a short discussion of the state of the literature on measures of mobility using cell phone data, including: what measures are out there, what has been done with them, if there have been any assessments of their performance, and how well they perform.

In general, there is only one measure of cell phone record derived mobility in the literature- the radius of gyration (ROG). There has been little assessment of its performance and it has rarely been used and assessed in models. Thus we know little about it's performance, accuracy, or efficacy. We include the ROG as one of our key measures in this paper and run it through our complete assessment with the other new measures we developed.

Data

Cell phone data

Our primary data of interest in this study is cell phone call data records (CDRs), which are permanently recorded by mobile phone service providers all over the world for billing purposes. A CDR is generated every time a user makes or receives a call or text message and contains the unique identifier of the caller and the callee; identifiers for the initial and final cellular antennae (towers) that handled the caller's call; the date and time the call was placed as well as the duration of the call. Coupled with a dataset describing the locations (latitude and longitude) of cellular towers, these massive datasets provide the approximate location of the caller when placing the call. Because cell phone owners generally make multiple calls during the course of even a single day and evening, CDRs can be used to study the movements, or mobility patterns, of individuals. Of course, the more calls a person makes, the more accurate will be our understanding of their actual mobility.

The CDR data we use in this study come from a single mobile phone provider in Rwanda and cover the years 2005-2009. During this time, cell phone penetration of the country increased dramatically, but was still likely under 50% by 2009. Our data from one mobile provider include about 1.5 million users by 2009, or about 15% of the entire population of the country.

These data are inherently selective in two ways. First, and less concerning, is that our data cover cell phone users from only one provider in the country. However, this company held a monopoly in the mobile phone market until 2008, when it was joined by another, and 2009 when a third company entered competition. Even during these last two years of our study (2008-2009), our cell phone records represent a large proportion of the Rwandan users and there is little reason to believe that there is significant selection in users between the three providers.

The second source of selection, which requires further discussion, is that our data cover only cell phone users, and exclude non-users. Research suggests that in many poorer countries, such as Rwanda, cell phone users are more likely to be wealthy or middle class, urban, and male (citations). As cell phones become more common, even in poor countries they are reaching almost 100% penetration, this selection problem decreases. However, for our study, it remains an important concern.

(Note: We will add more discussion here on exactly how this could affect our results and how our results should be understood in light of this selection issue.)

Contextual data

We also use a variety of spatially detailed contextual data in Bayesian models that test our mobility measures. Population density was derived from Oak Ridge National Laboratory's 2008 LandScan data[1]. LandScan is a unique and innovative population data set that provides ambient population estimates (average over 24 hours), based on disaggregation of census counts using spatial data and imagery. At approximately one kilometer resolution (30" x 30"), LandScan is one of the finest resolution global population distribution datasets available. Spatially explicit data on major roads, the capital Kigali, and international borders come from Open Street Map, a global open-source mapping project[2]. With these data, we created measures of distance to each feature.

Setting- Rwanda and the recent history of cell phone use and accessibility

Rwanda is a small country of approximately 10,000 square miles, which is about the size of the US state of Massachusetts. It is largely hilly at moderately high altitude with a temperate to sub-tropical climate. Rwanda generally ranks low on indices of human development, with an estimated 57% of the population poor and 38% extremely poor in 2006. The population density is amongst the highest in Africa, at 900 people per square mile and most Rwandans (87%) live in rural areas, as shown in Figure 1.

[Figure 1 about here- Population density map and location of cell towers.]

[1] http://web.ornl.gov/sci/landscan/
[2] www.openstreetmap.org

Despite the rather poor and rural conditions of the population and contentious history, the Rwandan economy and living conditions of the people have improved throughout the 2000's. Although the economy is primarily based on subsistence agriculture, coffee and tea have been successfully developed as cash crops for export and are now considered to be of excellent quality on world markets. Road networks have been improved. As shown in Figure 2, paved roads now connect the capital Kigali with most major cities, towns, and major international border crossings.

[Figure 2 about here- map of roads, Kigali, borders.]

Along with these changes, mobile phone technology also rapidly expanded in Rwanda throughout the 2000's. Starting in 2005, data from the mobile company show an average of over 1.8 million uses (calls or texts) per month. As shown in Figure 3, this consistently increased to over 15 million uses per month by the end of 2008. At the same time, and also shown in Figure 3, the number of active cell towers increased from 58 to 264 between January 2005 and May 2009. Figure 1 presents the distribution of all cell towers across the country that were in service at any time between these dates and makes clear that this expansion of towers vastly improved the accessibility of mobile technology in towns and rural areas throughout the country

[Figure 3 about here- cell phone users and cell towers.]

Measures of Mobility

To contribute to the study of mobility using cell phone data, we developed four new measures: Number of Towers Used, Maximum Distance Between Towers, Area of Convex Hull, and Perimeter of Convex Hull. We tested these four measures, along with the existing mobility measure, Radius of Gyration (ROG). All of these measures are calculated on a monthly basis. For example, Number of Towers Used refers here to the number of towers a person used within one month and is recalculated each month for each person. All measures could be calculated at different time frequencies as needed and appropriate. There are similarities and differences between measures in different dimensions, thus it is important to carefully examine and assess each measure and what it is actually measuring. We now describe each measure separately and illustrate each with a series of graphs and statistics.

Number of towers used

This measure (which we call NTU hereinafter) is a count of the number of cell towers that an individual used. The range of NTU goes from one to a maximum of 188 towers. Measurement of NTU is clearly contingent on tower density; the more towers in a certain area, the higher a person's mobility will be. In other words, a person living in Kigali, which has a high tower density, will have a higher NTU measurement than a rural person who moves the same distances during the month. However, there are a finite number of towers that could be used even in the densest area, so, past a certain threshold it is necessary to move around the entire country to obtain a high mobility score via the NTU measure. This is shown in Figure 4 on a map of the towers used by the most mobile person as measured by NTU.

[Figure 4 about here- maps of highest mobile person by each measure.]

As shown in Table 1, the person who was most mobile for the NTU measure, was also amongst the most mobile on most other measures as well. However, looking down the NTU column in Table 1, the NTU measure is somewhat volatile; the person who ranked highest on this measure (Person A) used 188 towers, whereas the highest mobile person for other measures (Persons B-E) used much lower counts of eight through 85 towers. This could suggest that NTU measures a different aspect of mobility than other measures, a possibility that we address further in subsequent sections of this study.

[Table 1 about here- table of highest mobile people, comparison of measures.]

Maximum Distance Between Towers

This measure (which we call MD hereinafter) measures the distance between the towers used in one month that are furthest from each other. As shown in Figure 4, the most mobile person by the MD measure used much fewer towers than the most mobile person by the NTU measure. Also different from NTU, the MD measure is clearly not affected by tower density; if a person lives in an area with high tower density, this will not affect the MD measure any more than for a person living in an area with low tower density. In other words, long distance travel is still necessary to rank highly on this measure.

Comparing MD to other measures, as shown in Table 1, the most mobile person as measured by MD is Person B with an MD measurement of 240. Person B is also amongst the most mobile for other measures. Looking down column 3 in this table, the high MD score of 240 for Person B is extremely close to the MD score for the highest mobile people under all other measures. Thus, it appears that MD is not a very volatile measure and may function similarly to some other measures presented here.

Area of Convex Hull

The measure of area of the convex hull (ACV) is created by calculating the area of the polygon that is created by the furthest towers that a person used. This is graphically presented in Figure 4. Similar to MD, this measure takes into account large distances between towers and does not account for the number of other towers that were used inside the polygon. However, it can take several towers into account and it is not heavily impacted by tower density.

In comparison to other measures, in Table 1, ACV produces fairly stable results. The most mobile person as measured by ACV also ranks highly on mobility via the other four measures. ACV is somewhat volatile; it ranges from a high of 26,327 for Person C to less than half that, 12,548 for person E. This suggests that ACV measures a different aspect of mobility than some other measures, particularly MD and RoG. Alternately, it appears similar to PCV which we discuss next.

Perimeter of Convex Hull

The measure of perimeter of the convex hull (PCV) is created by calculating the perimeter of the polygon that is created by the furthest towers that a person used. As shown in Figure 4, this is the same polygon used for the ACV measure and the PCV is thus similar in many ways to the ACV. Amongst similarities, this measure does not take into account the number of towers that are used inside the polygon and is not heavily impacted by tower density.

In comparison to other measures, as shown in Table 1, PCV not surprisingly behaves similarly to ACV when we compare the most mobile person across measures. Notably, the most mobile person via PCV is not the same as the most mobile person via ACV. Otherwise, this measure is relatively stable; the most mobile person via PCV also scores high on other measures.

Radius of Gyration

This measure (which we call RoG hereinafter) is calculated in several steps. First, a centroid is created between all the towers visited in the last month. Second, the distance of each tower from the centroid is calculated and squared. Finally, the squares are summed to create the RoG. This is graphically shown in Figure 4. This measure, which has been used in the literature but not well assessed until now, takes into account all towers visited. Similar to the NTU, it is affected by tower density; in other words, people who live in areas with high tower density will have higher RoG scores than people who live in areas of low tower density but the same movements throughout the month.

Comparing the RoG to other measures, we find it is somewhat different from most of them. The most mobile person via RoG had very low mobility under the NTU measure (only 8 towers visited throughout the month) and fairly low mobility under the ACV and PCV measures as well. In addition, there is some volatility whereby the most mobile person via RoG (Person E) had an RoG of 124, while the most mobile people under other measures (Persons A-D) had about half that, with RoGs of 60-69. This suggests that the RoG measures different aspects of mobility than all four of the other measures.

Other comparisons of mobility measures

Thus far, our comparisons of all five measures used the most mobile person under each measure; in other words, an outlier. Another useful comparison is to look at average mobility calculations for the whole population over time. As shown in Figure 5, despite the differences in measures discussed above, all five of them produce a similar pattern in average mobility over time of the cell phone using population of Rwanda.

[Figure 5 about here. Trends in mobility over time, for all 5 measures.]

A third useful comparison of measures is to examine which towers had the largest effect on mobility, as determined by each measure. Where two measures result in similar towers being important, we can suggest that these measures are similar in the aspects of mobility that they address. Alternately, where two measures result in different towers being important, we can suggest that these measures address different aspects of mobility. These graphical comparisons are shown in Figure 6.

[Figure 6 about here. Maps of most mobile towers for each measure.]

(Note: We will add more description of this comparison.)

Modeling Mobility

One of the primary purposes of creating these measures is to enable their use in models of mobility around the world. As such, we further interrogate our measures to examine how they perform in the modeling context. To do this, we devise a Bayesian linear model with two levels. At the first stage of the hierarchy we have a linear regression whose response variable is a mobility measure and whose explanatory variables are the indicators associated with each active cell phone tower. An indicator takes value 1 if a person visited the corresponding tower and 0 otherwise. At the second stage of the hierarchy we have a linear regression whose response variable is the vector of regression parameters associated with the indicator variables from the first level and whose explanatory variables are the contextual variables, including distance from the nearest major road, distance from Kigali, distance from an international border, and population density. The contextual variables describe key information related to the spatial location of each tower.

We apply the two stage linear model separately for each of the five mobility measures and for CDRs from each of the 51 months spanned by the Rwandan data (we excluded two months, May 2005 and February 2009, due to technical difficulties in recording accurate information which led to most of the CDRs being unrecorded). The number of samples n represents the number of individuals who made calls in the wireless provider's network in each month (ranges from about 150K to a little over 1M). The number of explanatory variables of the first level regression represents the number of towers that were active during that particular month (a tower is inactive if it did not handle any calls). The number of explanatory variables of the second level regression represents the number of contextual tower level variables. Note that both n and p1 vary from month to month, but p2 remains fixed for all months.

For each of the 5x51 = 255 datasets, we ran the Gibbs sampler which draws samples from the joint posterior distribution of model parameters for 50,000 iterations with a burn-in of 5; 000 iterations. The Gibbs sampler converged quickly since it involves direct draws from multivariate normal/inverse gamma distributions.

Modeling Results

Note: We will finish this section. The two main points will be:

1. Mobility measures perform better when geo-spatial contextual measures are controlled. This creates better model fit. Demonstrate with NTU measure and present maps that show importance of cell towers calculated in models with and without contextual measures. NTU performs better in models with contextual measures.
2. Mobility measures perform similarly and as expected when predicted by contextual measures. We will demonstrate this Figure 7, which presents a time series graph (2005-2009) of the effect of distance to major road on mobility. It includes a trend line (in black) and 95% confidence intervals (in grey shading). There is a negative relationship, meaning the further a person lived from a road, the less mobile they were. This is what could be theoretically expected, and thus demonstrates that the measure performs appropriately when associated with contextual data.

Conclusion

Note: We will finish this section.

References
To be added.

Tables and Figures

Figure 1. Map of Rwanda with population density and location of cell phone towers

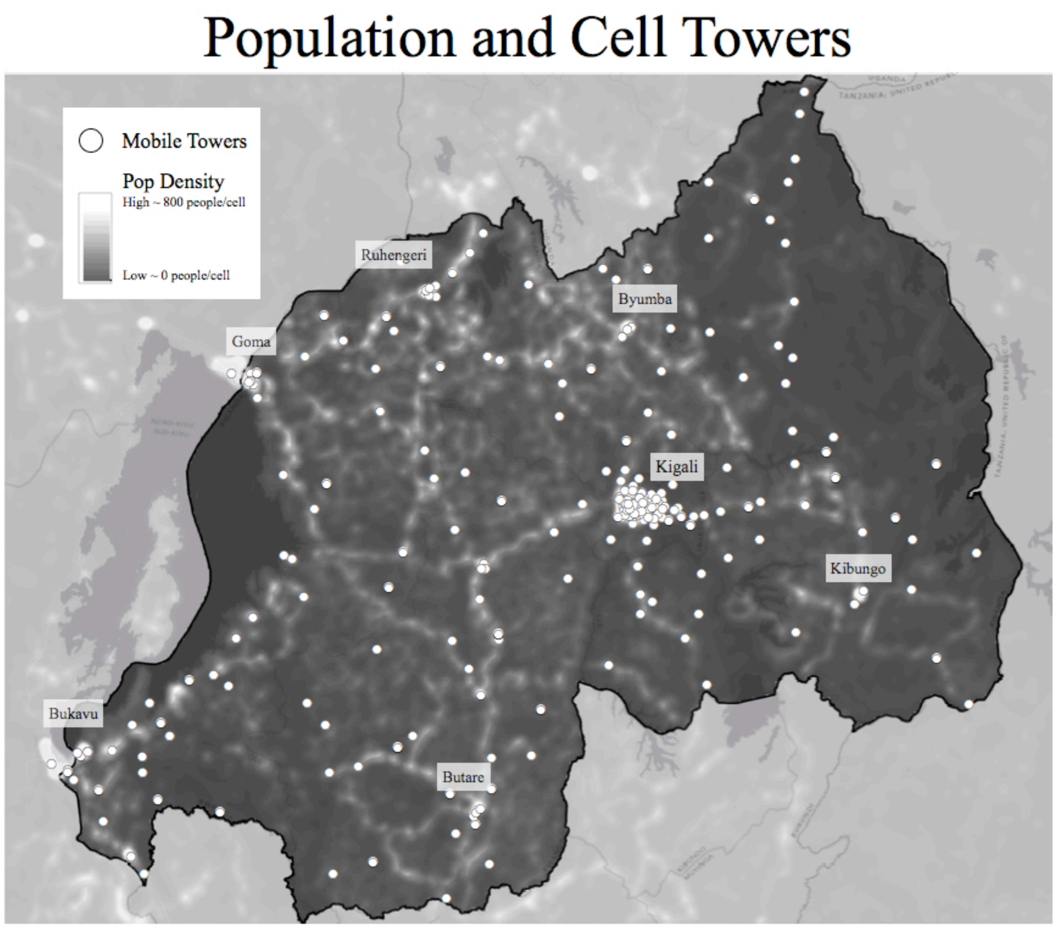

Figure 2. Map of Rwanda, with major roads, international borders, and Kigali.

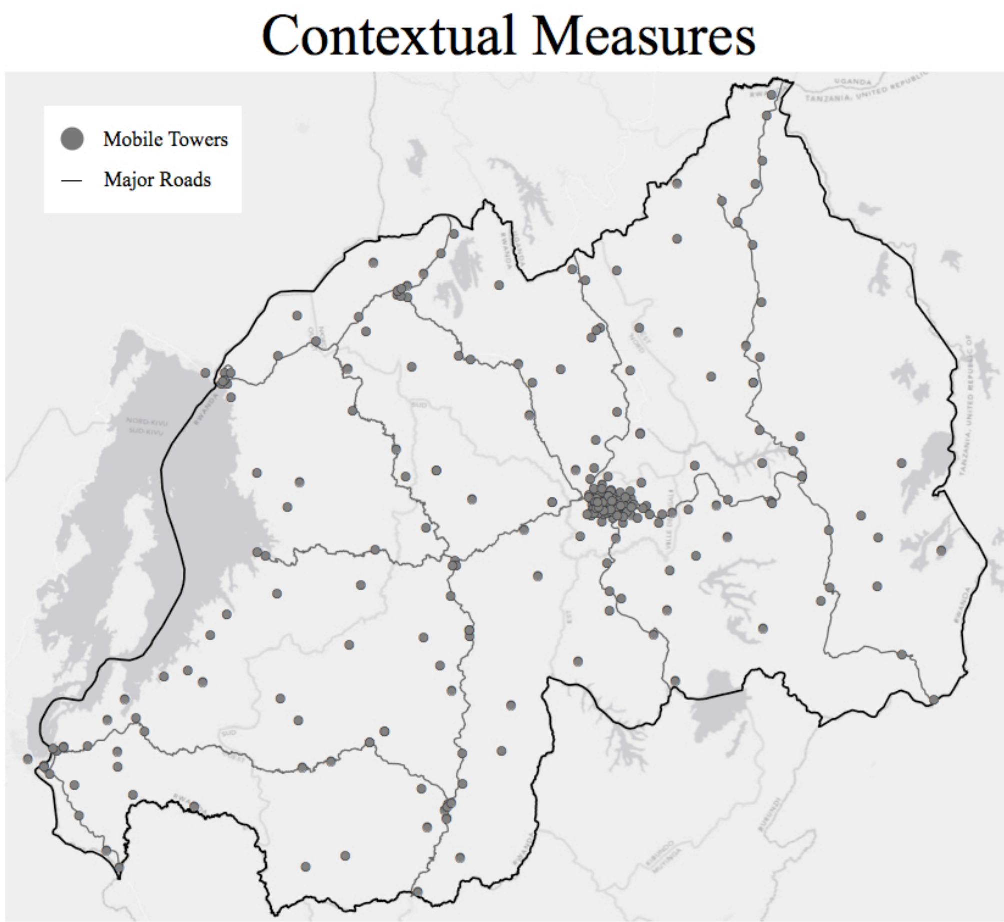

Figure 3. Trends in the number of cell towers and number of cell phone users in Rwanda, 2005-2009.

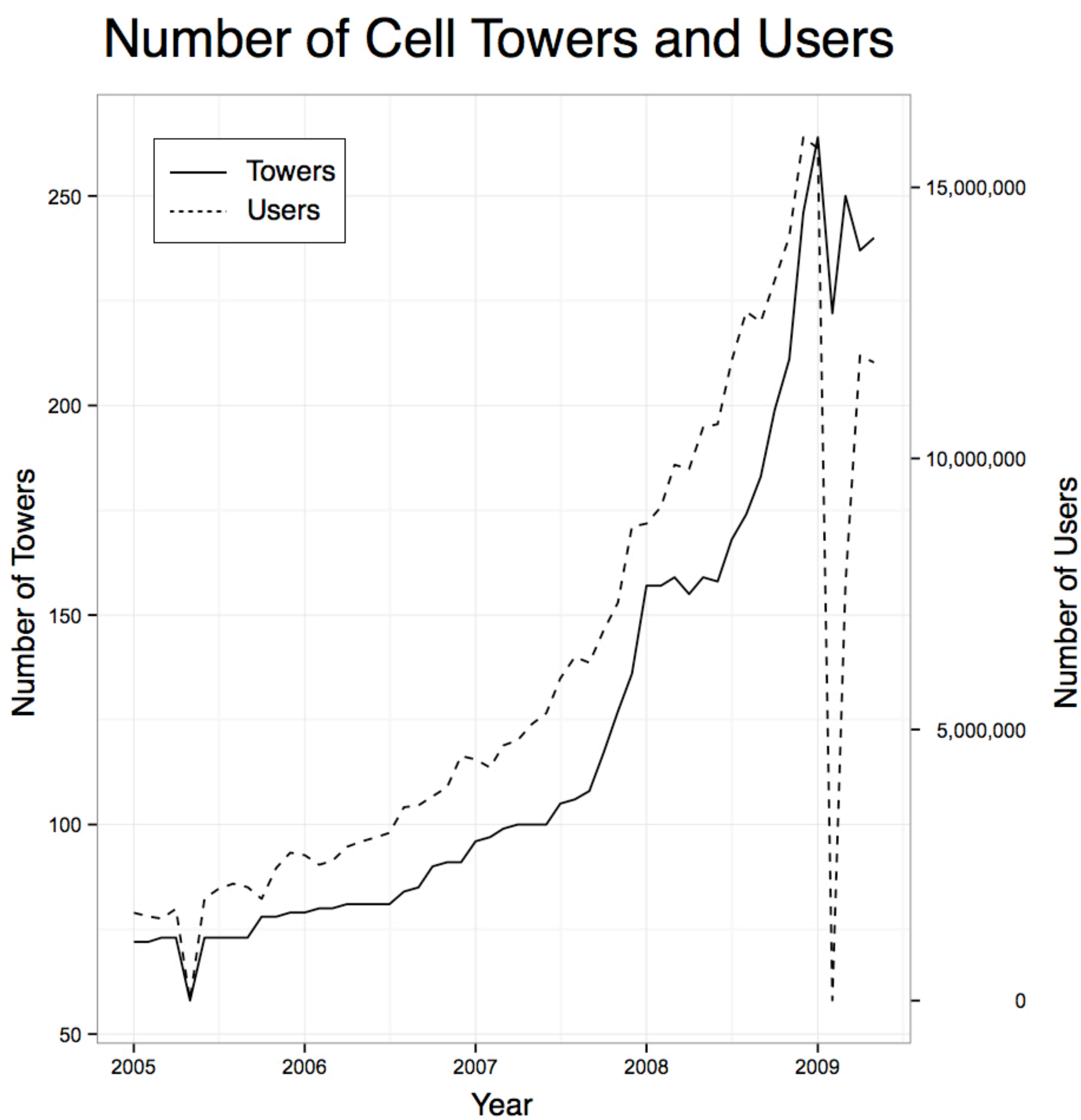

Figure 4. Maps of calculations for each mobility measure.

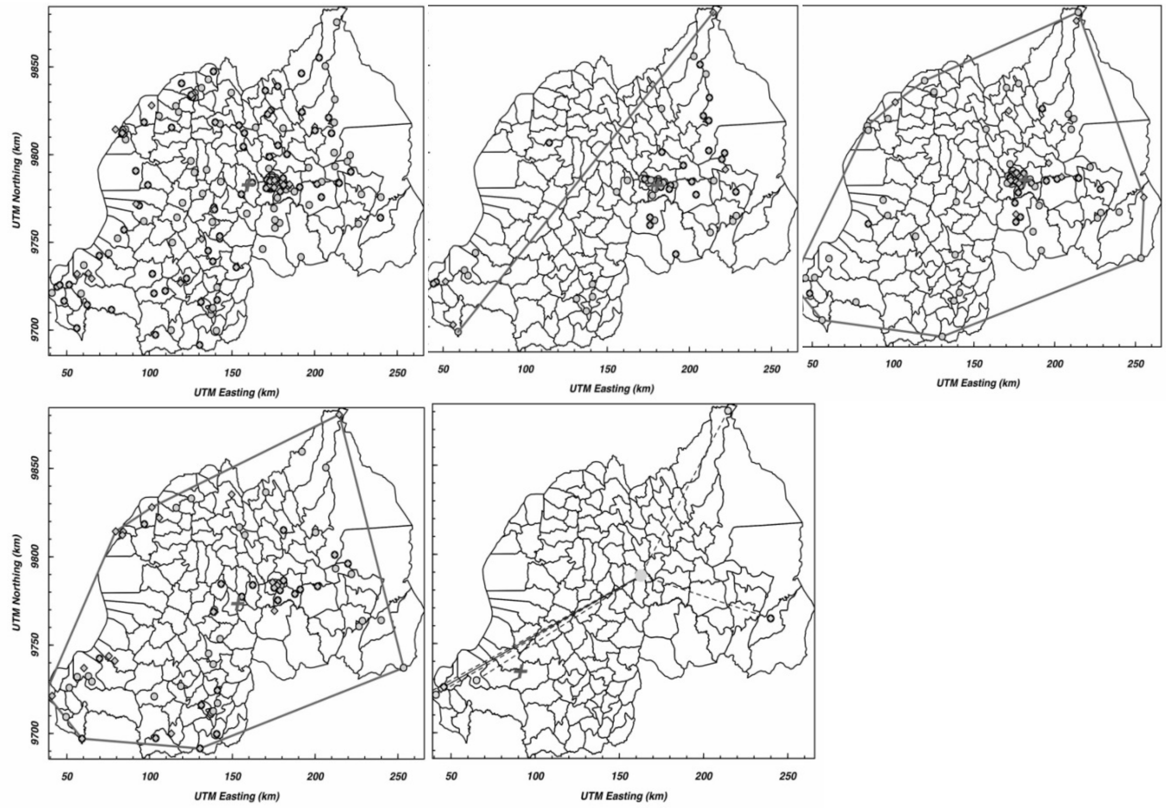

Figure 5. Trends in average mobility over time, compared for five mobility measures.

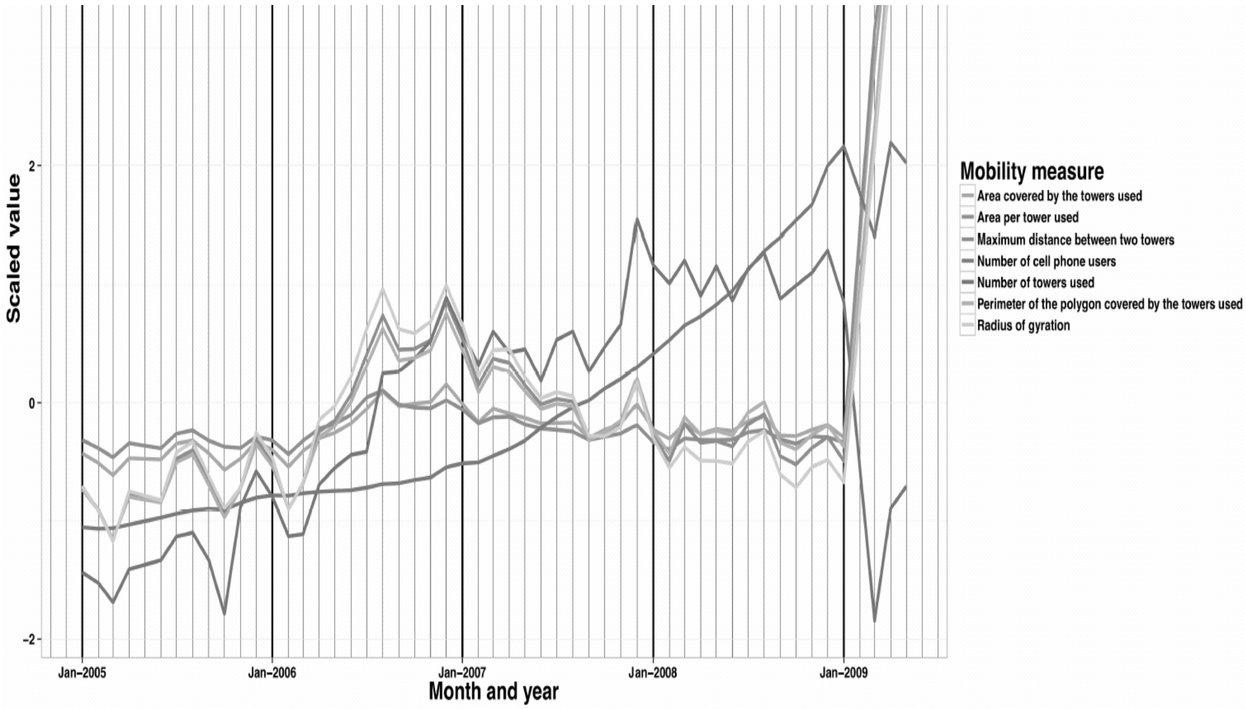

Figure 6. Most important cell towers for mobility, comparison of different mobility measures.

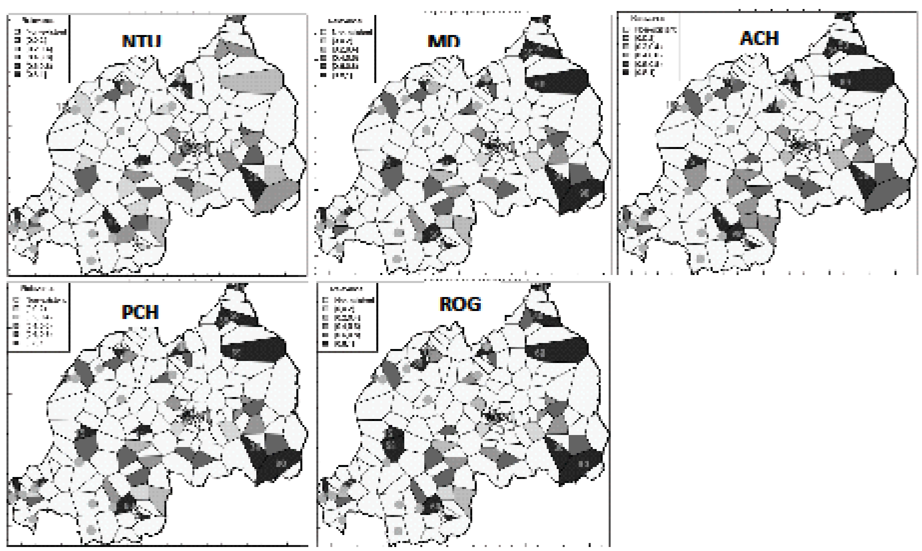

Figure 7. Plot of the effect of distance to the nearest major road on mobility, using NTU measure.

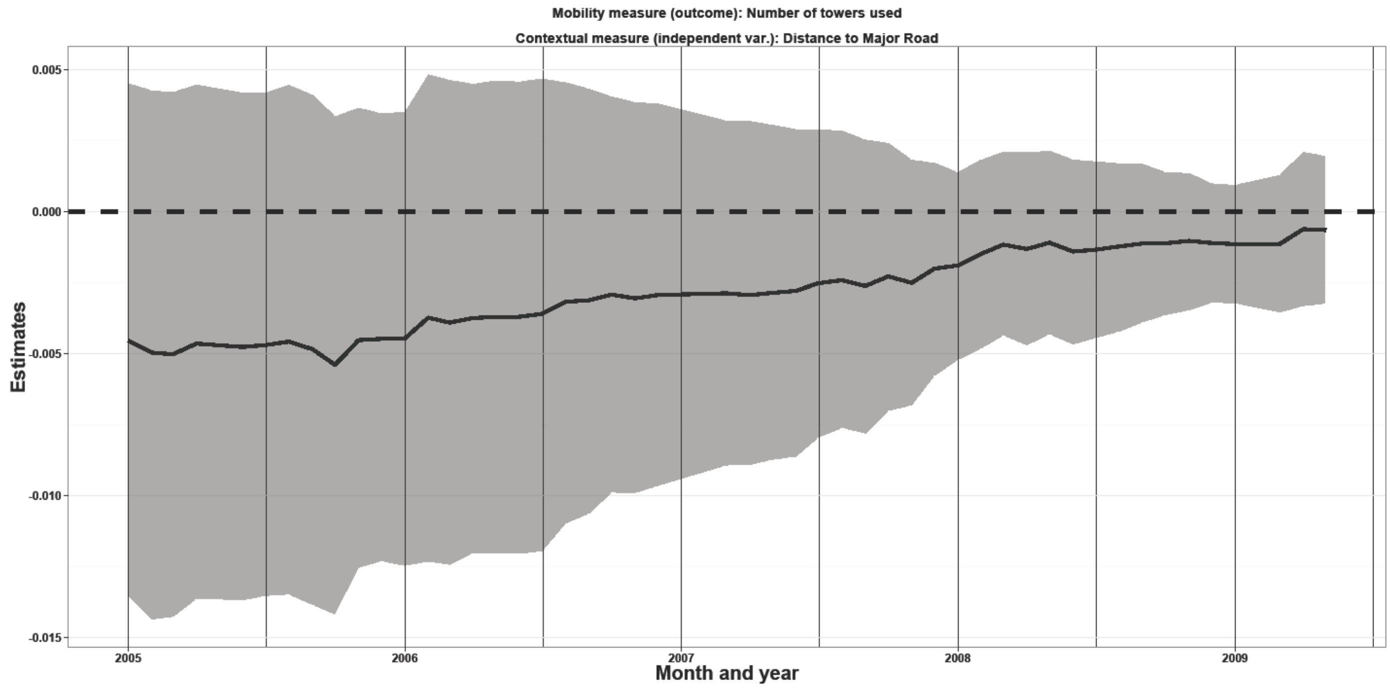

Table 1. Mobility scores for people who ranked most mobile under each measure.

	NTU	**MD**	**ACV**	**PCV**	**RoG**
Person A (most mobile person for NTU)	**188**	235	23,184	599	60
Person B (most mobile person for MD)	75	**240**	17,372	576	64
Person C (most mobile person for ACV)	85	239	**26,327**	636	63
Person D (most mobile person for PCV)	74	240	25,663	**636**	69
Person E (most mobile person for RoG)	8	240	12,548	564	**124**