Paper 1846-2014

Revealing Human Mobility Behavior and Predicting Amount of Trips Based on Mobile Data Records

Carlos Andre Reis Pinheiro, KU Leuven, Belgium

ABSTRACT

This paper reveals the human mobility behavior in the metropolitan area of Rio de Janeiro, Brazil. The base for this study is the mobile phone data provided by one of the largest mobile carriers in Brazil. Mobile phone data comprises a reasonable variety of information, including data about time and location for call activity throughout urban areas. This information might be used to build users' trajectories over the time, describing the major characteristics of the urban mobility within the city. A variety of distribution analyses is presented in this paper aiming clearly describes the most relevant characteristics of the overall mobility in the metropolitan area of Rio de Janeiro. In addition to that, methods from physics to describe trends in trips such as gravity and radiation models were computed and compared in terms of granularity of the geographic scales and also in relation to traditional data mining approach such as linear regressions. A brief comparison in terms of performance in predicting the amount of trips between pairs of locations is presented at the end.

In analytics terms, this human mobility study is a new approach to better understand consumers' behavior. The most often approach to perform such analysis is by using several profile information. How customers use products and services, when, in which frequency, recency and amount, how they pay; if they pay; how long they have been using and so forth. This set of information is quite individual and definitely describes the consumer's behavior in the company's relationship perspective. But, how customers relate to each other, how they affect to each other when using products and services? The first set of information describes who the customers are. The second one describes "how" they are. That question can be answered by using a social network analysis approach, describing how customers relate to each other.

INTRODUCTION AND THE PROBLEM STATEMENT

Human mobility analysis reveals relevant knowledge about subscribers behavior as well as urban planning, traffic forecasting or spread diseases (González et al 2008). Human mobility studies based on call detail records may also disclosure an approximation of the human motion within particular geographic areas such as metropolitan and urban areas, big cities, entire states and even countries, including mobility behavior and migration trends (Simini et al 2012).

This particular human mobility study was conducted by using mobile phone data from one of the largest telecommunications company in Brazil. The overall analyses were performed upon six months of call detail records, revealing the average pattern of behavior for users travelling throughout the city over the time. The total amount of records handled in this study is about 3.1 billion records, comprising 2.1 million subscribers, handed out through 5,332 towers cells.

The analysis of the human mobility in urban areas may arises relevant knowledge about the population's behavior in terms of motion and displacements (Rubio et al 2013), allowing companies and government organizations in better planning traffic and public transportation, optimizing telecommunications' networks, deeper understanding the vectors of spread diseases, accurately defining procedures in case of unexpected scenarios such as disasters and catastrophes, and preparing big events such as World Cup and Olympic Games, as is the case of the city of Rio de Janeiro.

Many researches have been carried on human mobility studies in developed countries (Liu et al 2013). However, much less is known about the urban mobility and human motion in developing countries such as in Brazil. In this study we aim to cover a set of distribution analyses in order to describe the mobility patterns in a big metropolitan area like the city of Rio de Janeiro. By describing the overall human mobility pattern in Rio de Janeiro, it is possible to compare the outcomes from developed and developing cities and therefore to analyze possible differences in trends of human displacements.

Another goal in this paper is to apply approaches from physics such as gravitation and radiation models to better understand mobility trends and predict trips between locations. In addition to the physics' models, a set of data mining models based on regressions were also applied and the outcomes were compared in terms of trips prediction between locations.

Understanding human mobility behavior, trends in amount of trips between locations, and overall displacements within big cities, such as Rio de Janeiro, is a quite good foundation to plan traffic routes, improve public transportation systems and network communications, particular toward to big events forthcoming such as World Cup in 2014 and the Olympic Games in 2016.

METHODOLOGY AND DATA PREPARATION

Mobile phone data provides a set of information about caller and called numbers, assigning them to tower cells spread out through urban areas. These tower cells hand out the calls made by the subscribers, and give us some useful information about their geographic locations in specific points of time. The range of information mapped for the tower cells is particular for each mobile carrier, providing basically data about latitude and longitude, radius, and address description such as neighborhood, district, city, county, and state. In this study, we are taking into account just the information about the subscribers for the mobile carrier which has provided the mobile phone data. Upon this data, we have the tower cell when the subscribers make or receive calls, indicating the geographic position they were at the particular time of the call. Even though it is a sort of approximation, recent studies (Candia et al, 2008) shows that by using the appropriate techniques, mobile phone data may offer the possibility of characterizing statistically the human trajectories and travels in a urban area scale.

Over the day, subscribers can make and receive calls in different periods of time, but not always when they are moving. Also, the distribution of the call activity follows a sort of a power law, where most of the subscribers make and receive a few amount of calls over the time and a few subscribers make and receive a significant number of calls. Also, mobile phone records are sparse over the day and hold indefinite gaps in space and time. For this reason, we decide to monthly aggregate the call activity in order to assembly the travel history for the subscribers. By doing this, calls received and made in different days, keep just the call time information, building therefore a possible trajectory over the day. There is a differentiation between weekday and weekend in doing this aggregation. Thus, calls made and received from Monday to Friday are aggregated by time period slots of one hour, as well as calls made and received on Saturday and Sunday. Several studies (Schneider, 2013) (Jiang, 2013) show that most of the individuals have a recurrent travel activity, visiting in most of the times the same subset of locations, and in very few times different locations than the usual ones. Therefore, we believe that the method to reduce the sparsity of the calls and the spatio gaps by merging days keeping the time period still reveals the average subscribers' behavior in terms of urban mobility.

Upon this approach, if a particular subscriber makes a call on Tuesday at 1 pm at location a, then another call on Wednesday at 10 am at location b, and receives a call on Thursday at 4pm at location c and finally receives another call on Friday at 6 pm at location d, then a possible trajectory for this subscriber would be the path b – a – c – d. Considering the assumption of the recurrent trajectories and frequent displacements, we may assume that this particular subscriber is most likely to be at location b around 10 am, at location a around 1 pm, at location c around 4 pm and at location d around 6 pm. This assumption is important because as a matter of fact, subscribers don't make and receive calls anytime they move. We believe that this approach to aggregate mobile phone data consists in a reasonable and accurate method to diminish the sparsity and the spatio gaps within the data in building frequent trajectories and performing human mobility analysis over the time.

Hence, all six months were aggregated considering this aggregation approach, discarding the day and keeping the time period, splitting then the data aggregation process into two distinct datasets, containing data for weekdays and weekends. For further evaluation, we kept the weekday allowing us to perform distinct behavioral analyses for weekdays and weekends in terms of urban mobility.

In building the trajectories for all subscribers, some cleansing constraints were put in place in order to make the mobility data even more realistic, even though increasing a bit its sparsity. All calls were fitted into slots of one hour. A call at 1:32 pm then is flagged as a call in the time period 1 (between 1 and 2 pm), and so on. Locations visited just one time over the timeframe were discarded from the analysis. These locations might represent an unusual behavior of the subscribers in traveling throughout urban areas. Also, once the information about call activity was summed, the most frequent visited location was considered for each time slot. For instance, a particular subscriber s may visit location a 10 times during the timeframe analyzed at the time period 1, location b 5 times at the same time period, and location c just 2 times. Once location a is the most visited at the time period 1, we assume that this location is the most likely position to find subscriber s between 1 and 2 pm. Finally, once we are concern about the subscribers' motion, if a particular subscriber has the same most frequent location in consecutive time slots we keep just the first one. For example, if subscriber s has location a as the most frequently visited in time slot 10, 11 and 12, just the combination of location a and time slot 10 is kept. The others time slots are discarded.

Also, the start time for the displacement activity during the day was shifted in this study. Based on the call activity distribution by period of times, we assume that the day starts at 5 am in this mobility analysis and ends at 4:59 on the day after. Figure 1 shows the call activity by time.

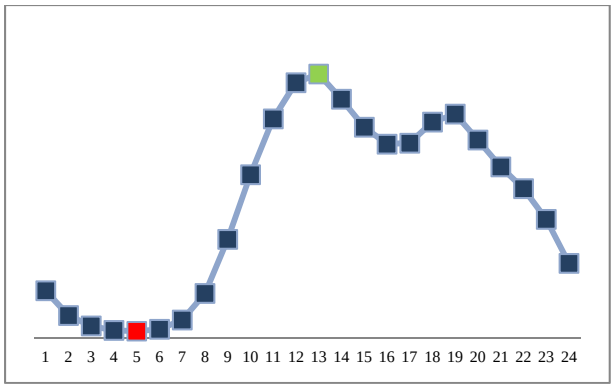

Figure 1: Call activity distribution by time. The minimum level of activity is between 5 and 6 am, and the day was considering by starting at this time period. The call activity increases onward and reaches the peak between 1 and 2 pm. Then starts decreasing until 4 pm and resumes increasing from 5 pm until 7 pm, when the activity decreases until 5 am on the day after.

In order to complete the cycle of the trajectories, we also calculated the presumed domiciles and workplaces for each subscriber. The presumed domicile is the most frequent location visited between 7 pm and 7 am on the day after. Analogously, the presumed workplace is the most visited cell during the period between 9 am and 5 pm. The periods of 7 am and 9 am and from 5 pm and 7 pm were considered as commuting periods.

The presumed domiciles in our study play an important role in building the trajectories and then in analyzing the human mobility patterns throughout the city of Rio de Janeiro. In order validate the method to achieve the presumed domiciles we compared the population found based on our method and the last census available for the city of Rio de Janeiro. The presumed domiciles explain over the 82% of the observed population in the city according to censored data, as following in the figure 2.

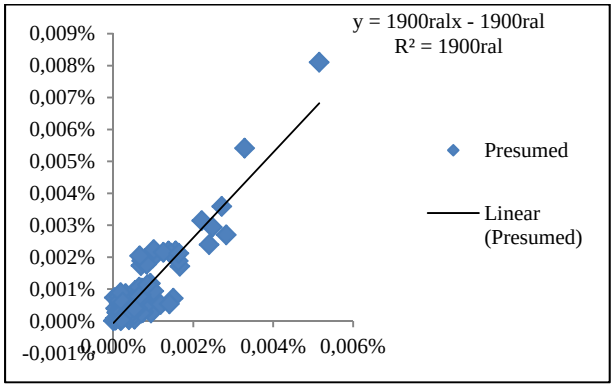

Figure 2: Presumed domiciles computed by the method explain over than 82% of the censored data observed.

The most visited locations for both presumed domiciles and workplaces has a minimum threshold of 3 visits over the timeframe. Figure 3 shows the distribution of the presumed domiciles and workplaces computed in this study.

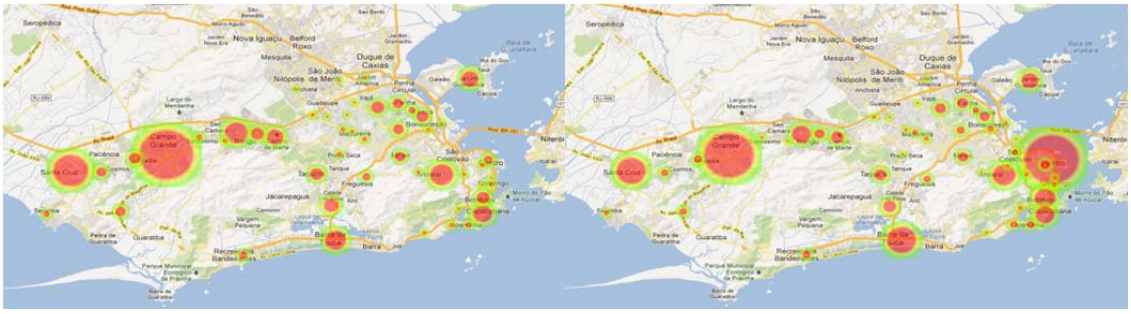

Figure 3: Presumed domiciles and workplaces calculated according to the most frequent visited cells over specific time periods. The rank of neighborhoods in terms of population is very similar to the data census provided by government agencies.

Then, the theoretical first displacement for the subscriber's trajectory is from the presumed domicile to the location at he or she performs the first call activity. Similarly, the last displacement is from the location assigned to the last call activity to the presumed domicile. The displacements between presumed domiciles and presumed workplaces were used to analyze the frequent commuting displacements throughout the city. Figure 4 shows the most relevant displacements in terms of presumed domiciles and workplaces.

Figure 4: The first picture shows the most relevant displacements between presumed domiciles and workplaces in the city of Rio de Janeiro. The second picture shows the frequent destinations from the most populated neighborhood based on the presume domicile. The third picture shows the frequent origins to the most populated presumed workplace.

It is important to notice that the displacements start shifted during the weekends in relation to the weekdays, as expected. People wake up and start moving earlier during the weekdays to work and a bit later during the weekends. Figure 5 shows that difference.

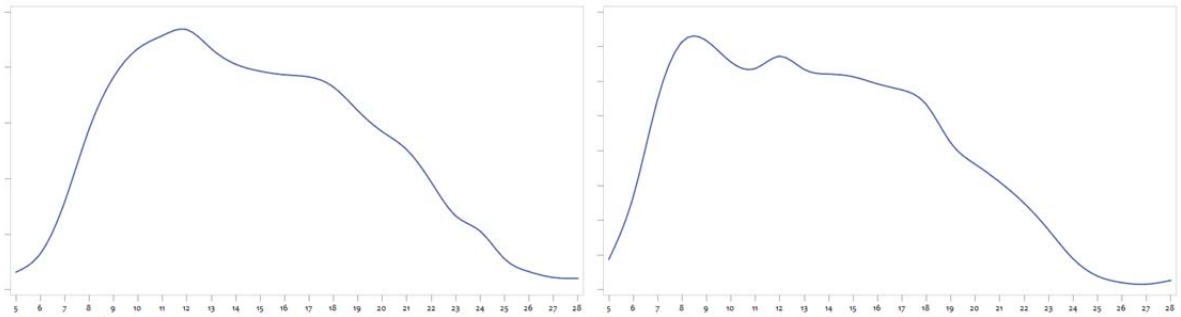

Figure 5: The first picture shows the first displacements by time during the weekdays, and the second one shows the first displacements during the weekends. It is possible to notice that during weekday people start moving earlier than during weekend.

TRAJECTORIES MATRIX AND BEHAVIORAL ANALYSIS

The mobile phone data, considering all top frequent cells visited by the subscribers, in all time period slots, is then transformed into a matrix containing the spatio information. For each subscriber, this matrix contains all time periods

slots, and for each one of it, the most frequent cells visited throughout the timeframe considered. Upon this matrix, it is possible to build a set of analyses about the human motion over the time and space. It is a tridimensional matrix comprising the subscriber, the time period and the location. Figure 6 shows a graphical representation about how this matrix is shaped.

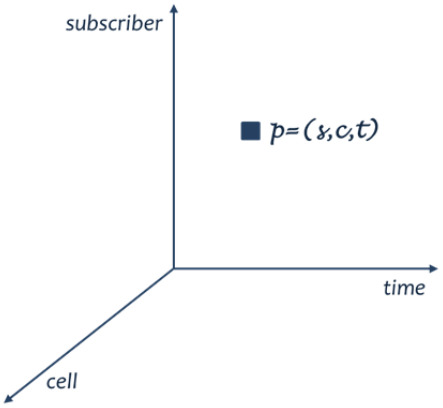

Figure 6: The matrix of trajectories contains the location of each subscriber in all time period slots over the timeframe. Based on this tridimensional matrix is possible to analyze the trajectories travelled by all users over the day.

This matrix basically consists in store all sequences of distinct locations a particular user visited in two different points in time, represented by the following formula:

,

where is the location i of the user a in time t, and where is the location j of the user a in time t+1.

It is quite important to notice that all analyses presented onwards are based on this matrix. There is a significant distinction in analyzing the regular datasets – comprising just the call detail records for instance – and this matrix. The matrix proposed presents just the subscribers' motion, and therefore, all analyses upon it are assigned to the users' mobility instead of the users' activity indeed. Suppose that a particular subscriber makes 10 calls a day, but in the same location, being handled by a single tower cell. This subscriber definitely has a call activity but there is no motion assigned to it. Another subscriber, for example, makes just 2 calls, but in distinct locations, being handled by different tower cells. In spite of a less call activity in relation to this subscriber against the other one, there is a clear displacement event configured in here, when the second subscriber moves from the first tower cell to the second one.

Base on this approach, a call activity in reality is a displacement activity, and all displacements – a pair of distinct locations visited – is a vector of movement (Park et al 2010). Figure 7 shows the average focuses of displacements in the city of Rio de Janeiro, as well as the major vectors of movement within the metropolitan area over the time.

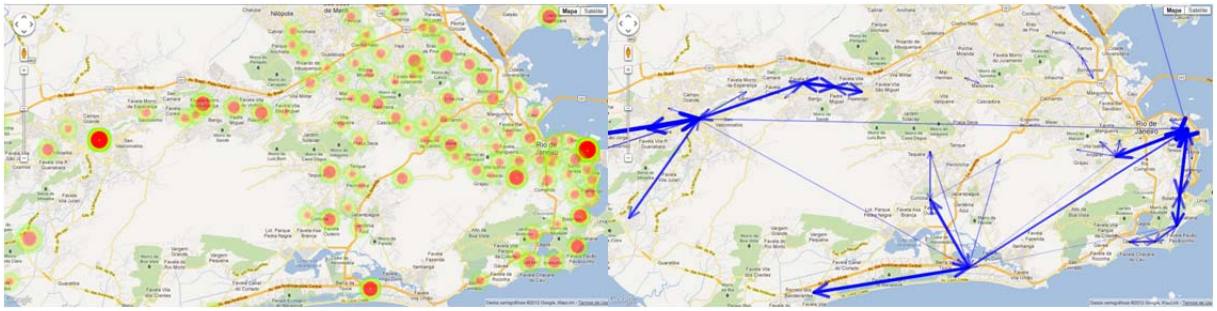

Figure 7: The displacements activities and the major vector of movements in the city of Rio de Janeiro over the time.

PATTERNS IN HUMAN MOBILITY IN THE CITY OF RIO DE JANEIRO

An overall analysis about the human mobility behavior in the city of Rio de Janeiro starts by looking at the call activity, which demonstrates indirectly, according to the objective of this study, the average pattern in human mobility in terms how period of time. Figure 8 shows how subscribers move over the days and hours. All days, whether business days or weekends, have two peaks, around noon and at the end of afternoon, when people move around most frequently. The peak in activity takes place on Thursdays, particularly at 6 and 5 pm and then noon and 11 am. Wednesdays are the second peak in activity, followed by Tuesdays, Fridays, Mondays, Saturdays and finally Sundays.

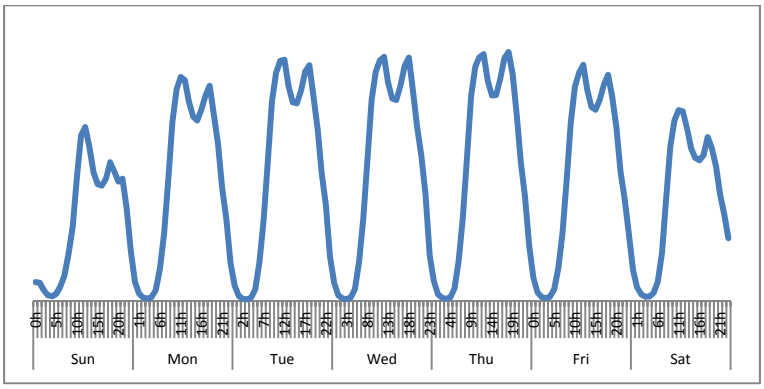

Figure 8: The call activity by weekday and time period over the timeframe of analysis. Displacements' activity has two peaks every day, around noon and at late afternoon. The highest activity takes place on Thursdays.

In human mobility, one of the major characteristic of the overall displacements throughout the cities is that some locations are more visited than others. This fact highly contributes to the human motion over geographic areas, and it is a crucial attribute in traffic and public transportation planning, in vector analysis for spread diseases, in procedures definition in the case of disasters and catastrophes, among many others business and social applications. Figure 9 shows how the locations in the city of Rio de Janeiro are frequently visited by mobile users. A log distribution is also presented.

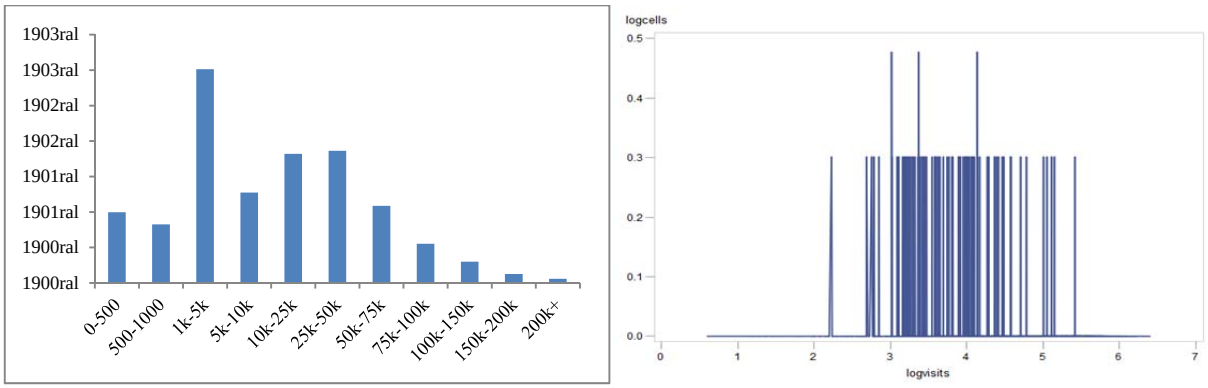

Figure 9: The amount of visits for each cell over the city of Rio de Janeiro.

The visits' distribution is not linear, where some locations are more visited than others. For instance, 50% of the tower cells receive less than 10k visits monthly, while 10% of the tower cells receive a range of visitors between 70k and 400k. Figure 10 shows how distributed is the amount of visitors on a monthly basis.

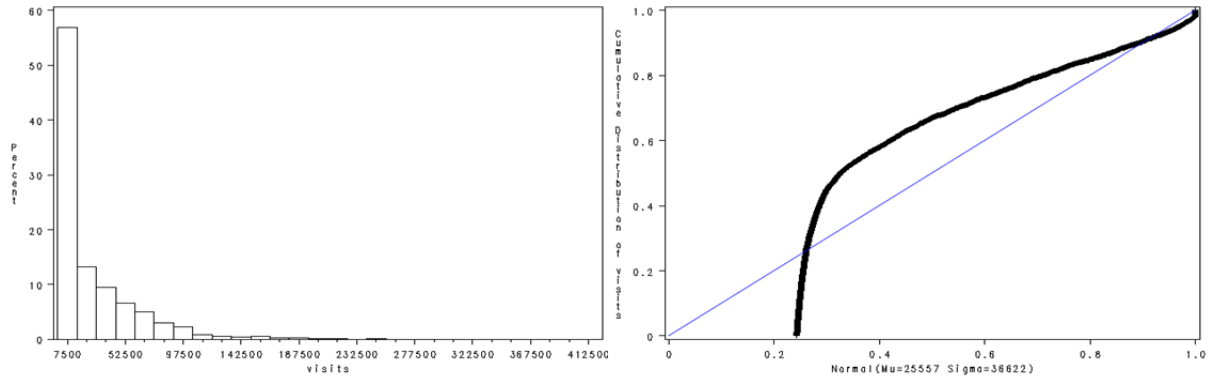

Figure 10: The amount of tower cells' visitors on a monthly basis in the city of Rio de Janeiro.

Communications patterns are well known to be highly heterogeneous, sometimes following a power law, where some users rarely use mobile phone while other subscribers make a substantial amount of calls quite frequently. The call activity distribution in the city of Rio de Janeiro is similar to the distribution in other metropolitan areas, as previously presented in (Jiang, 2013).

Also, the distribution of the call activity by subscribers is close to others metropolitan cities. The amount of calls per user is presented in figure 11, as well as the log distribution of it.

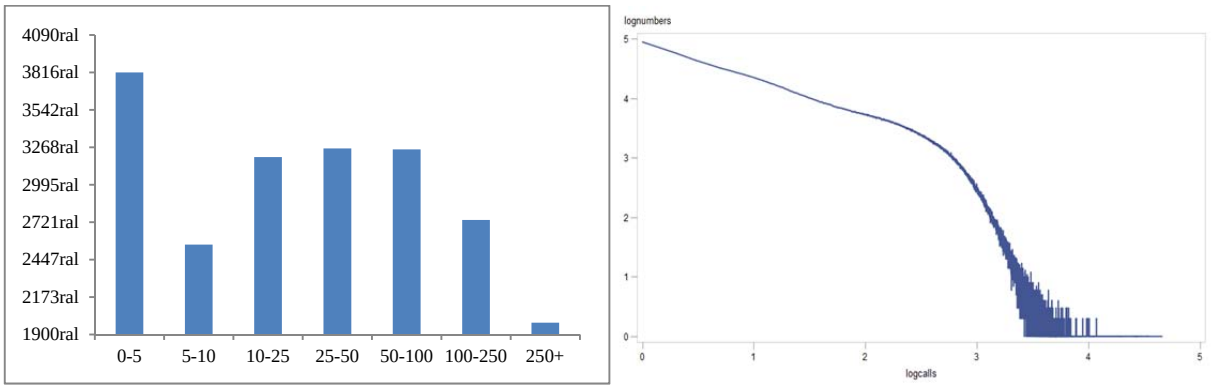

Figure 11: The call activity by subscribers. Most of the users have low activity and some of them have high call frequency over the time.

The mobility activity for the users follows a kind of power law, where 10% of the users move just one single time, 15% of the users move between 1 and 5 times, 25% between 5 and 24 times, another 25% of the users move between 24 and 61 times, and just 10% of the users move more than 100 times on a monthly basis. It is important to reinforce in here that the call activity in this study is measure assigned to displacements. If users make consecutive calls at the same location, there is no displacement, and therefore, there is no call activity. Figure 12 shows the distribution of the displacement activity for the subscribers.

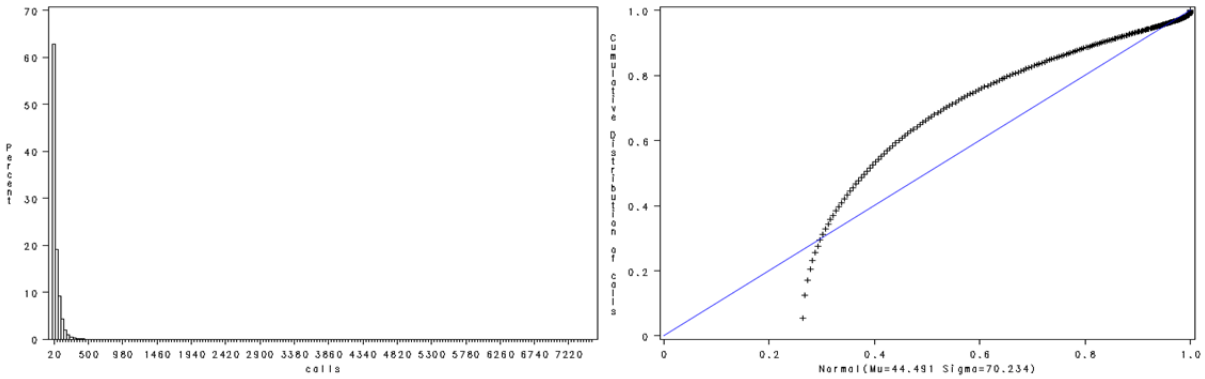

Figure 12: The amount of displacements for the subscribers on a monthly basis in the city of Rio de Janeiro.

Another important measure is the degree of mobility, expressed by the amount of distinct cells visited by each subscriber over the time. Analogously to the call activity – displacement activity in this study, some users visited few locations as per usual, and others may visit several geographic locations over the time, indicating short trips and long trajectories in terms of frequent displacements, respectively. Figure 13 shows the amount of distinct cells visited by subscribers and the log distribution of it.

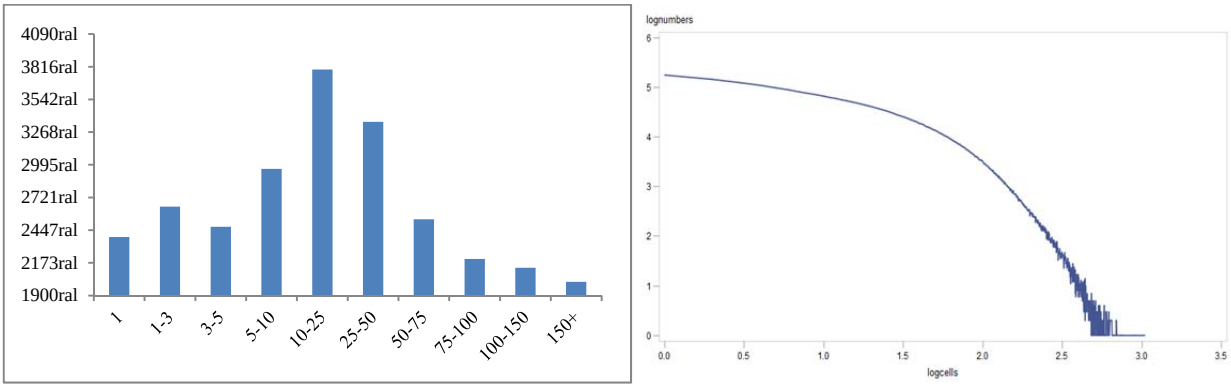

Figure 13: The amount of distinct cells visited by subscribers over the time.

The average amount of mobility in the city of Rio de Janeiro may be considered as quite high. Most of the users visit frequently a reasonable amount of distinct locations during their average trips. The pattern of mobility may be described as follows: just 5% of the users frequently visited a single location over the time, and another 5% visit 2 distinct locations, while 25% of the users visit between 6 and 16 distinct locations, and another 25% of users visit between 16 and 38 distinct locations throughout their frequent trips. Also, 10% of the users visit more than 70 distinct locations during their frequent displacements, reaching up to 200 distinct geographic areas. Figure 14 shows the distribution of the distinct locations visited by the users over the time.

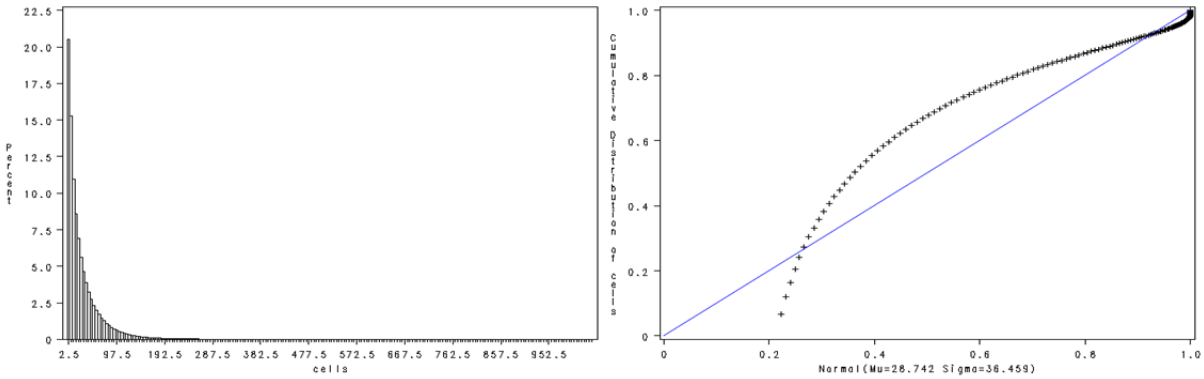

Figure 14: The amount of distinct locations visited by the users throughout their frequent displacements in the city of Rio de Janeiro.

Finally, another descriptive measure for the pattern of mobility is the average distance traveled by the users. Rio de Janeiro is the second largest city in Brazil, the third largest metropolitan area in South America, sixth in Americas and twenty-sixth in the world. It is then expected that people have to travel quite long distances in moving around the city, both for work and leisure. Figure 15 shows the average distance traveled by the users and the log distribution of it.

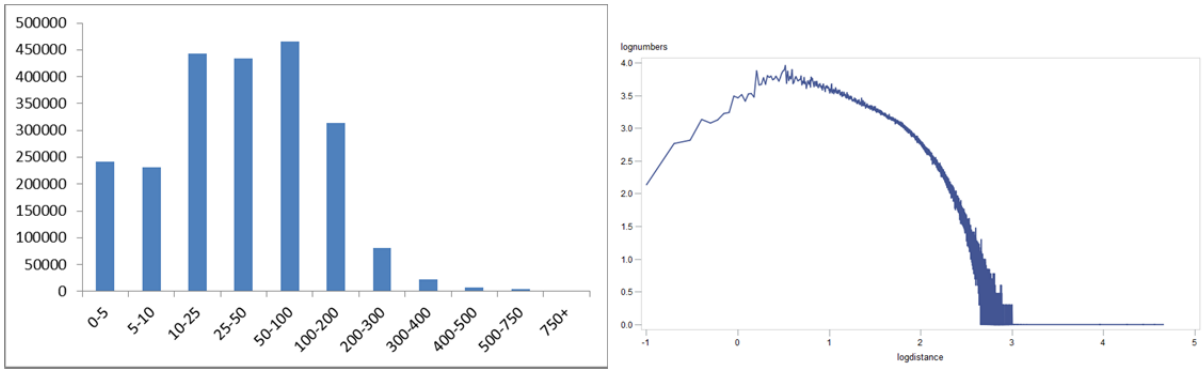

Figure 15: The average distance traveled by the users over the time.

Looking at the displacements in the city of Rio de Janeiro, just 10% of the users travel in average less than 5 km considering their frequent trajectory. As a big metropolitan area, 25% of the users travel between 15 and 35 km, another 25% travel between 35 and 80 km and 10% of them travel more than 150 km, reaching a significant amount of 350 km in their frequent displacements over the time. Figure 16 shows the amount of distance traveled by the users.

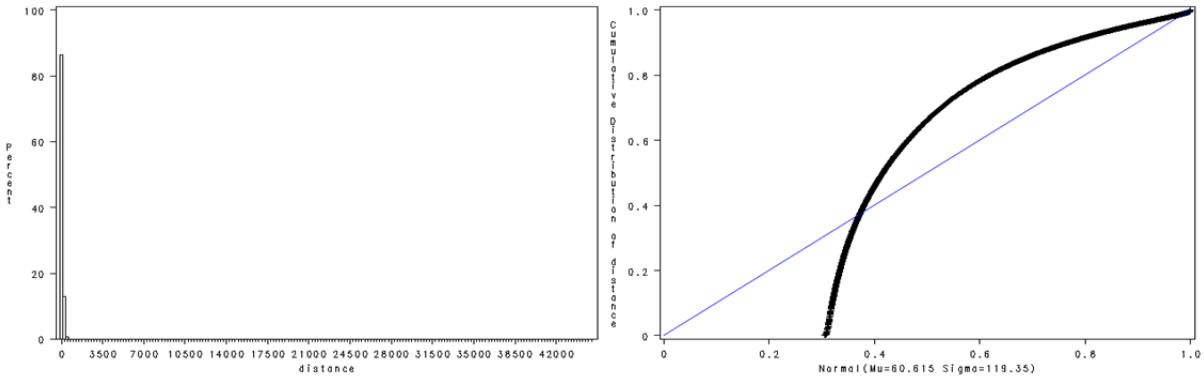

Figure 16: The amount of distance traveled by the users throughout their frequent displacements in the city of Rio de Janeiro over the time.

In the next sections we are going to cover the predictability of the amount of trips between a pair of locations based on physics models such as the gravity and the radiation models, as well as upon statistical models such as regressions.

In order to perform this evaluation, we are going to consider a subset of trips within the state of Rio de Janeiro, taking into account just the trips departing or arriving within the great metropolitan area. From now on all analyses will consider this subset of data when referring to trips in the city of Rio de Janeiro, comprising a substantial amount of 3,042 tower cells, 135 neighborhoods and 15,816,882 possible distinct trips. All predictive estimative are based on these possible trips.

GRAVITY AND RADIATION MODELS FOR PREDICTING TRIPS BETWEEN LOCATIONS

In the recent years, gravity and radiation models provided a method to raise new insights and useful knowledge about population movements modeling (Yan et al, 2013). These models stand on a solid foundation and they can accurately reproduce mobility patterns varying from migration paths to commutes trajectories.

The gravity model usually requires some parameters to define constraints on the generation and attractions of flow such as distance and cost of travel. It states that the commuting between two locations is proportional to the product of the population of these two places and inversely proportional to a power law of the distance between these two places. On the other hand, the radiation model requires just the spatial distribution population as input, and no further adjustable parameters. It basically depends only on the origin population, the destination population, and on the population in a circle is the origin and the radius is the distance between the origin and the destination (Masucci et al, 2012).

The lack of adjustable parameters in the radiation model makes easier to the model the mobility patterns upon mobile phone records such as we have in this study. In spite of the radiation model might be not applicable to predict human mobility at the city scales, there are some evidences that this type of model which is quite successful in reproducing mobile patterns at large spatial scale can raises reasonable mobility trends in great metropolitan areas.

The gravity model assumes that individuals are attracted to other locations as a function of the distance between these two places associated to the cost of the travel distance between them. The radiation model assumes that individuals are attracted by the nearest locations rather than the farther locations with more opportunities due to the hypothesis of limited resources of mobility and high cost of travels. In both models the travel distance is a crucial factor in decision making of users in commuting between two locations. So, the distance distribution is an important statistical attribute to describe the human mobility behavior.

The easiest gravity model is expressed by a simple function of the populations in origin and destinations locations by the distance between them, using a scaling factor, as presented in the following formula.

$$T_{ij} = k \frac{P_i P_j}{d_{ij}},$$

where T_{ij} is the number of trips between locations i and j, P_i is the population in location i, P_j is the population in location j, and d_{ij} is the distance between locations i and j.

Also, the simplest radiation model is expressed as a function of the population in origin and destination locations by the population in a circle whose center is the origin and the radius is the distance between origin and destination (minus populations in origin and destinations locations). The formula is presented as following.

$$T_{ij} = T_i \frac{P_i P_j}{(P_i + P_{ij})(P_i + P_j + P_{ij})},$$

where T_{ij} is the number of trips between locations i and j, P_i is the population in location i, P_j is the population in location j, and P_{ij} is the total population in the circle of radius r_{ij} (the distance between locations i and j) centered at location i (excluding populations of locations i and j).

The gravity and the radiation models were employed in the urban area in the city of Rio de Janeiro upon six months of history in terms of human trips, considering the amount of 2,879 tower cells and 31,483 pairs of locations – possible trips throughout the city. As stated before, this is a subset of the entire data considering just the trips departing or arriving in the downtown of the city.

These models assumed a straight correlation between the amount of trips and some attributes in the mobility pattern, such as the distance between the pair of locations, the population in the origin location, the population in the destination location, the population in the area of radius equals the distance between origin and destination locations

and center in origin location –excluding the origin and destination populations – e finally the amount of trips started in the origin location.

The correlation between the amount of trips and these variables are shown in picture 17 in a log x log distribution graph.

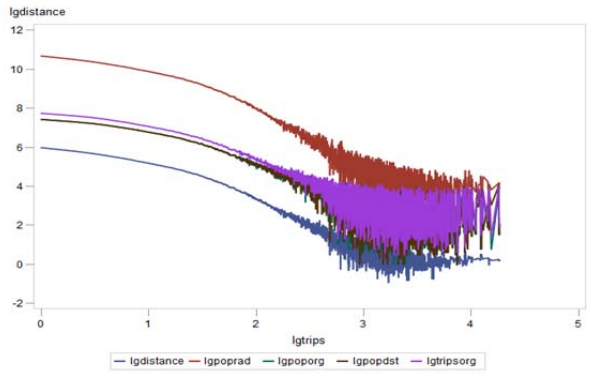

Figure 17: The correlation between the distance traveled by the users and the population in the origin and destination locations, in the circle of the origin location, and the amount of trips started on the origin location.

This graph basically shows that for short trips the correlation of those variables is not much tight as it is for long trips. It might be explained by the granularity of tower cells in the geographic location in the city. There are some tower cells covering short spaces for many people and some tower cells covering long spaces for few people. The tower cells are established according to the population density and therefore, some very short trips can take place in our study. The distances assigned to the trips, the populations in origin, destination and radius, as well as the amount of trips in the origin might be overlapped in that scenarios and the correlation gets lower. For long trips these variables are better established in space and the correlation between trips and they suit better.

The gravity model performed better than the radiation model in this particular study, reaching a coefficient of determination R2 of 0.5788. R2 is a measure about the goodness of fit, and it shows how well the predictive model approximates to the observed data.

The gravity model takes into account the populations in origin and destination, as well as the distance between origin and destination in order to predict the amount of trips between these two locations. On the other hand, the radiation model takes into account the population of origin and destination locations – similarly as the gravity model – but also the amount of trips started in the origin and the population of the area with center in origin and radius of the distance between origin and destination (excluding origin and destination populations). As explained before, upon low granularity such as tower cells, short trips may have an overlap in distances and population, and probably due to this, the performance of the radiation model was lower than the gravity in this particular case. The radiation model has reached a coefficient of determination R2 of 0.3675. Figure 18 presents the performance of the gravity and radiation models in terms of variation of the predicted amount of trips and the observed trips.

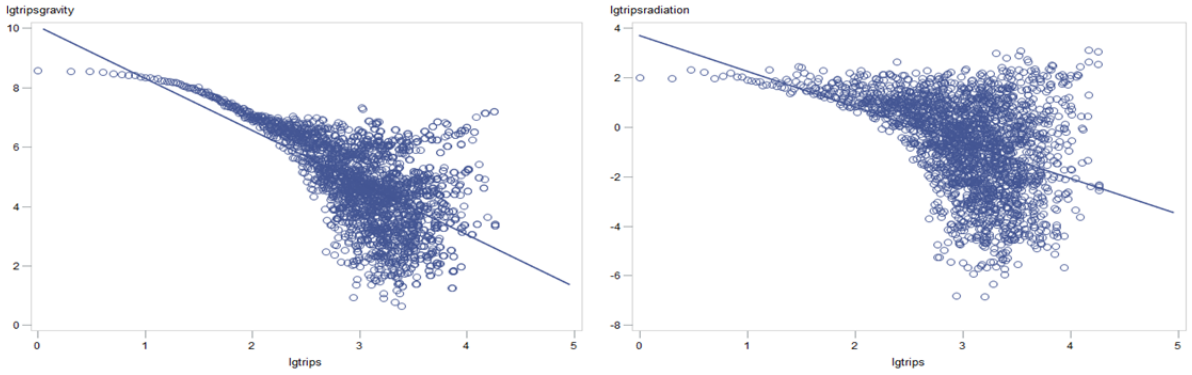

Figure 18: The performance of the gravity model in predicting the amount of trips between tower cells in the city of Rio de Janeiro reached a R2 of 0.5788. The radiation model on the other hand has reached a R2 of 0.3675.

STATISTICAL MODELS TO PREDICT THE AMOUNT OF TRIPS BETWEEN A PAIR OF LOCATIONS

In the previous section we noticed how the gravity and radiation models performed in predicting the amount of trips between tower cells in the city of Rio de Janeiro. The gravity models explains about 58% of the trips in the city and the radiation models explains less than 37% of the trips considering the granularity of tower cells.

Now we are going to employ a set of statistical models mostly based on linear regression in order to predict the amount of trips between tower cells. Aiming to have a fair comparison between the types of models, we are going to use the same set of attributes to predict the amount of trips, consisting in the population of the origin and destination locations, the distance between the origin and the destination, the amount of trips started in the origin location and the population in the area with center in the origin locations and radius of the distance between the origin and the destination locations.

The first model is a simple linear regression where the response variable Y can be predicted by a linear function of a regression variable X, including estimated intercept and slope , as presented in the following formula.

,

The second model is based on quantile regressions, which generalizes the concept of a univariate quantile to a conditional quantile given one or more covariates. This method models the effect of covariates on the conditional quantiles of a response variable by means of quantile regression. Ordinary least squares regression models the relationships between the covariates X and the conditional mean of the response variable Y given X = x. The quantile regression extends the regression model to conditional quantiles of the response variable such as the median or the 90th percentile.

The third model is a linear regression which detects and considers the outliers and provides resistant and stable results even in the presence of these unexpected observations. This robust regression limits the influence of the presence of outliers by using a high breakdown value method based on the least trimmed squares estimation. This breakdown value is a measure of the proportion of contamination than an estimation method can withstand and still maintain its robustness.

Figure 19 presents the performance of the linear regressions applied to the trips' data in the city of Rio de Janeiro. The first method, the simple linear regression reached a R2 of 0.7796, the second method based on quantile regression reached almost the same value of R2, 0.7792, and the third method based on robust regression reached a slightly better value of R2, 0.8201. These results mean that the simple linear regression and the quantile regression can explain over than 77% of the trips between tower cells in the city of Rio de Janeiro, while the robust regression explains over the 82% of the trips between locations within the great metropolitan area in Rio de Janeiro.

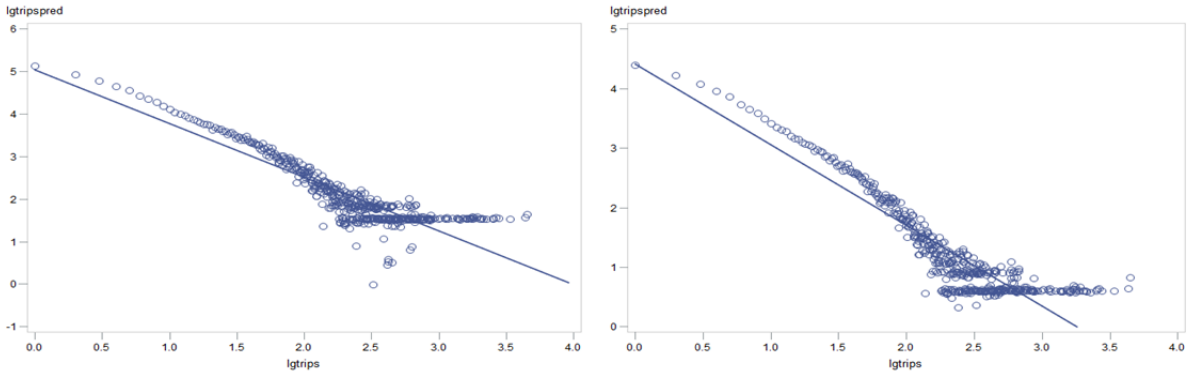

Figure 19: The performance of the simple linear regression and the robust regression in predicting the amount of trips between tower cells in the city of Rio de Janeiro reached a R2 of 0.7796 and 0.8201 respectively. The graph containing the performance of the quantile regression was suppressed once it has reached a very close performance than the simple linear regression model.

CONCLUSIONS

Human mobility behavior is a hot topic in science nowadays. This subject comprises lots of possibilities in better

understanding populations' behavior in terms of mobility, displacements and frequent trajectories upon cities, states and countries (Lee et al, 2009). One of the major sources of information about human mobility is phone data records from telecommunications carriers. As long as we gather more and more data from those mobile operators, and considering the mobile penetration within population in big cities, we are closer to approximate the subscribers' mobility behavior to the populations' mobility behavior. As more data collected from carriers, more closely we are to the real human displacements' behavior over the time.

Upon mobile phone records, scientists are able to aggregate raw information about space and time of subscribers' activities and then perform a set of human mobility studies. There are some approaches based on physics theories to predict the amount of trips between pairs of locations. These locations in some studies might be tower cells, or neighborhoods, or cities, or perhaps even countries (Yan et al, 2013). Upon high granularities such as cities and countries these studies usually are referred as migration behavior instead of mobility or displacement behaviors.

In this paper we performed the traditional models based on physics theories, such as gravity and radiation models. Both models in particular circumstances can explain and estimate quite well the amount of trips between pairs of locations, particularly in high scale space such as cities and countries. Very often, when using low geographic scaling the outcomes are not that good in terms of prediction of the amount of trips between locations. Indeed, we have experienced this in our study. Many researches have been showed this scaling issue, mostly in Europe, Asia and North America.

In order to compare the performance of the gravity and radiation models to statistical modes, mostly based on regressions, we perform all analyses upon tower cells. The gravity model had a medium performance, reaching an R-square of 0.5788. The radiation model had a poor performance, reaching an R-square of 0.3675. On the other hand, the regressions models had achieved reasonable performances, reaching 0.7796, 0.7792 and 0.8201 by using simple linear regression, quantile regression and robust regression, respectively. The performance of the best statistical models (robust regression) was twice as much as the best physics model (gravity model).

The statistical models; in addition to the machine learning models such as decision tree and artificial neural networks; might be an alternative approach to estimate the amount of trips between locations when the geographic scaling is low, like tower cells or small administrative areas within neighborhoods and cities.

REFERENCES

Candia, J., González, M., Wang, P., Schoenharl, T., Madey, G., Barabasi, A-L.2008. "Uncovering individual and collective human dynamics from mobile phone records." Journal of Physics A: Mathematical and Theoretical, Vol. 41 N. 224015.

González, M., Hidalgo, C., Barabási, A. 2008. "Understanding individual human mobility patterns." Nature, Vol. 453: 779-782.

Jiang, S., Fiore, G., Yang, Y., Ferreira, J., Frazzoli,E., González, M. 2013. "A review of urban computing for mobile phone traces: current methods, challenges and opportunities." Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

Lee, A., Chen, Y-A., Ip, W-C. 2009. "Mining frequent trajectories patterns in spatial-temporal databases." Information Sciences, Vol. 179: 2218-2231.

Liu, F., Janssens, D., Wets, G., Cools, M. 2013. "Annotating mobile phone location data with activity purposes using machine learning algorithms." Expert Systems with Applications, Vol. 40, Issue 8: 3299-3311.

Masucci, A., Serras, J., Johanson, A., Batty, M. 2012. "Gravity vs radiation model: on the importance of scale and heterogeneity in commuting flows." arXiv:1206.5735.

Park, J., Lee, D., González, M. 2010. "The eigenmode analysis of human motion." Journal of Statistical Mechanics: Theory and Experiment Vol. 2010.

Rubio, A., Sanchez, A., Martinez, E. 2013. "Adaptive non-parametric identification of dense areas using cell phone records for urban analysis." Engineering Applications of Artificial Intelligence, Vol. 26: 551-563.

Sarle, S. 1983. "Cubic Clustering Criterion." SAS Technical Report A-108, SAS Institute Inc.

Schneider, C., Belik, V., Couronné, T., Smoreda, Z., González, M. 2013. "Unraveling daily human mobility motifs." Journal of The Royal Society Interface, vol. 10 no. 84 20130246.

Simini, F., González, M., Maritan, A., Barabási, A-L. 2012. "A universal model for mobility and migration patterns." Nature, Vol. 484: 96–100.

Yan, X-Y., Zhao, C., Fan, Y., Di, Z., Wang, W-X. 2013. “Universal predictability of mobility patterns in cities.” Physics and Society, arXiv:1307.7502.

Ward, H. 1963. “Hierarchical grouping to optimize an objective function.” Journal of the American Statistical Association, 58, 236–244.

RECOMMENDED READING

- Pinheiro, Carlos. *Social Network Analysis in Telecommunications*. Wiley and SAS Business Series

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Carlos Andre Reis Pinheiro
Enterprise: Katholieke Universiteit Leuven
E-mail: carlos.pinheiro@kuleuven.be