Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast

Mirco Nanni[1(✉)], Roberto Trasarti[1], Barbara Furletti[1], Lorenzo Gabrielli[1], Peter Van Der Mede[2], Joost De Bruijn[2], Erik De Romph[2], and Gerard Bruil[2]

[1] KDD Lab, Isti CNR, Pisa, Italy
{mirco.nanni,roberto.trasarti,barbara.furletti,
lorenzo.gabrielli}@isti.cnr.it
[2] Goudappel Groep, Deventer, The Netherlands
{pvdmede,jdbruijn,edromph,gbruil}@goudappel.nl

Abstract. In this work we present an analysis process that exploits mobile phone transaction (trajectory) data to infer a transport demand model for the territory under monitoring. In particular, long-term analysis of individual call traces are performed to reconstruct systematic movements, and to infer an origin-destination matrix. We will show a case study on Ivory Coast, with emphasis on its major urbanization Abidjan. The case study includes the exploitation of the inferred mobility demand model in the construction of a transport model that projects the demand onto the transportation network (obtained from open data), and thus allows an understanding of current and future infrastructure requirements of the country.

1 Introduction

Population growth, massive urbanization and, particularly, the extensive increase of car use in the last century have led to serious spatial, transport, infrastructural and environmental problems in almost all urbanized areas. As a consequence, since the 1960s urban and transport planning methodologies were developed to forecast future traffic volumes and the expected use of infrastructure and facilities. The purpose of such forecasts is evident: infrastructure and urban planning provide keys to the mitigation and preclusion of transport and environmental problems. Today, urban and transport planning are major tasks of all public authorities.

The availability of spatial data on demographics, labor and land use has until now been a prerequisite for establishing an Origin and Destination matrix (OD matrix) for a transport model. It is a time consuming activity to obtain the necessary data in many developed countries. Moreover, in most developing countries the overall availability of data is very limited and, therefore, the use of transport models has never been a promising option for such countries. In this work we explore the possibility of deriving a proper OD matrix from mobile phone data, by using publicly available (free) transport network data and standard transportation modeling software, in order to build a basic transport demand model.

J. Nin and D. Villatoro (Eds.): CitiSens 2013, LNAI 8313, pp. 15–25, 2014.
DOI: 10.1007/978-3-319-04178-0_2,

In this way, we also provide evidence that also for countries or cities where many data seem to be lacking, now transport demand models can be created. This will on the one hand allow national and local authorities to have a far better understanding of transportation needs and challenges, and will help funding agencies and investors to better assess their potential risks and benefits.

In the next sections we will describe step by step the process we propose to derive transport demand models from phone data, and use it to build a (as far as we know) first transport demand model for Ivory Coast and its major urbanization Abidjan.

2 Background

This work tries to combine data mining of GSM traces with transportation modeling methodologies to gain insights into mobility in a monitored area, to allow what-if analysis through simulation or modeling.

GSM data have already been used to describe mobility in several studies, essentially based on the fact that a sequence of geo-referenced calls of users constitutes approximate trajectories of their movements. The key limitations of GSM data are that locations are only approximations and that sampling rate may be low and erratic. Works like [1] try to overcome these issues by working at a large geographical scale and/or under specific conditions (in that case, users where tourists in a large area). In the present work we follow a different approach, and try to exploit the relatively long temporal extension of a dataset to infer more reliable movement information. In particular, an approach similar to [2] (translated from GPS to GSM data) [6] and [9] is adopted, where we try to extract regular movements that repeat consistently in time, which therefore are less likely to be artifacts of the data sampling procedure, and use them to measure systematic mobility in the area (details are provided in next sections). Also, concepts like most favored location, which are exploited in this work, have already been applied in the scientific studies, e.g. [3], but mainly for simple distributions of a population or the recognition of specific activities, such as working, being at home, or leisure.

Macroscopic transport modeling methodology is well established [7,8]. This methodology is mainly implemented through commercial and academic software tools (e.g. OmniTRANS, Visum, Cube, Emme/2, TransCAD), and readily available. These tools can be used only by professionals with accurate knowledge of traffic theory and transport modeling experience. Macroscopic modeling has been used widely by governments and engineering firms to predict future transport network problems and for infrastructure planning. The current paper does not address network or transport planning as such. Its main purpose is to use data mining of GSM traces as an input for transport models, thus integrating these traces as widely and readily available sources of information into the transport planning realm.

3 Introduction of the Case Study

The data used in this work is composed of anonymized Call Detail Records (CDR) of mobile phone calls and SMS exchanges between five million of Oranges customers in Ivory Coast (corresponding to around one quarter of the national population) between December 1, 2011 and April 28, 2012. The data was made available by Orange in the context of the D4D (Data for Development) data challenge [10]. The data contains 10 samples taken in different time windows, each covering 50,000 individuals, corresponding to around 1 % of the population of customers. The IDs of individuals are changed from sample to sample and it is not known whether the sub-populations described in the different samples overlap, making it impossible to link data among different samples. The data provided contains for each observation the coordinates of the antenna serving the user during a communication, in other words the device is operating in an area covered by that antenna.

In general the coverage of an antenna is influenced by several factors: strength of the signal, the height of the pole, the orientation, the weather, the nearby buildings, etc. Since the provided data does not contain this information and it is not easy to retrieve it from external sources, we apply a well-known methodology in order to estimate the coverage of the antennas using only their spatial location. The method is called as centroid Voronoi tessellation and assumes that the space is partitioned into separate areas, each defined as the set of locations that are closer to our antenna than any other one. The partitioning of the space obtained will be used in all the following analysis.

4 Systematic Traffic Analysis and Transport Modeling

The basic events we are interested to spot in the data are systematic trips. Following the approach in [2], we define systematic trips as routine movements that users perform (almost) every day at (approximately) the same hours. By combining together the systematic movements of each individual in our population, we can obtain an estimated OD matrix that describes the expected flow of people between pairs of spatial locations as in [4].

Since the current GSM data are not detailed enough to detect whether a user stopped at a location or initiated a trip within an input sequence, we tackled the problem through a two-step procedure: first, we identified locations that are significant for the mobility of the individual, also called attractors; second, we identified movements between significant locations that occur with a high frequency, which are later aggregated across the whole population to fill in an OD matrix. The first step is performed according to the standard approach, also illustrated in [3]: the location where the largest number of calls took place is identified and labeled as L1 (most frequent location). Then, the second most frequent location is identified and labeled as L2. It seems likely that in most cases L1 corresponds to the home location and L2 to work or any other main activity of the individual, or vice versa. The second step is performed over the

sequences of L1 and L2 that appear in the traces of each single user. We checked the frequency of movements L1 L2 and L2 L1 within specified time slots. Each movement identifies a trip, and if its frequency is high enough, we assume it to be a systematic trip that the user performs during a typical day.

4.1 Detecting Users Attractors

The available data provide the information of the zone from which a phone call is started. Thanks to the large amount of data provided by the telephone operator it is possible to use the spatio-temporal footprint left by the users for the purpose of monitoring their movements in the territory.

Several studies [5] assert that most people spend most of their time at a few locations, and the most important ones may be labelled as *home* and *work*. In this section we will explain the methodology used for the extraction of such important locations, which we will call L1 (most important one) and L2 (second most important one) using the frequency of calls made by users. The location L1 relates to the antenna from which the user made the greatest number of phone calls. For both technical and infrastructure reasons due to the load balancing of the antenna, it may happen that the serving antenna for different calls made at the same place, may be different, even though such antennas are usually close to each other. To mitigate this effect, we have redefined L1 as a bigger area that also includes the adjacent cells. In Fig. 1, the most frequent location is represented by the central pink cell, but according to the method just described, we define as L1 the bigger area that includes the adjacent blue cells.

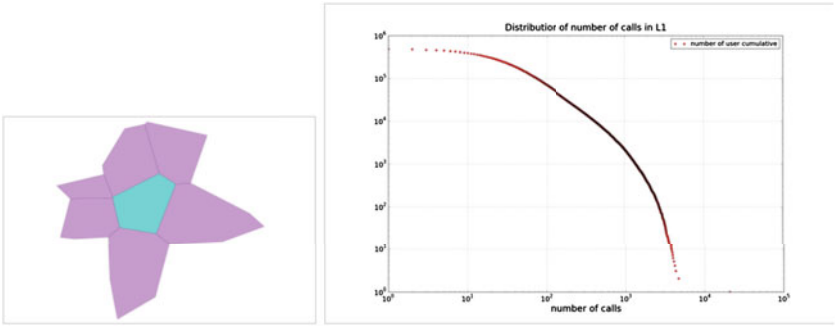

Fig. 1. Example of L1 detection (left), distribution of number of calls in L1 (right)

It is necessary to point out that, for the most of the users, the call frequency associated to L1 is quite low , thus rising issues of statistical significance and reliability of such a location (see Fig. 1). If we consider as a minimum frequency threshold for L1 of one call per day (therefore 15 calls in 15 days) from the area, only for 20 % of users (100 K) the associated L1 would result meaningful. The rest of this work assumes to use such subset of locations.

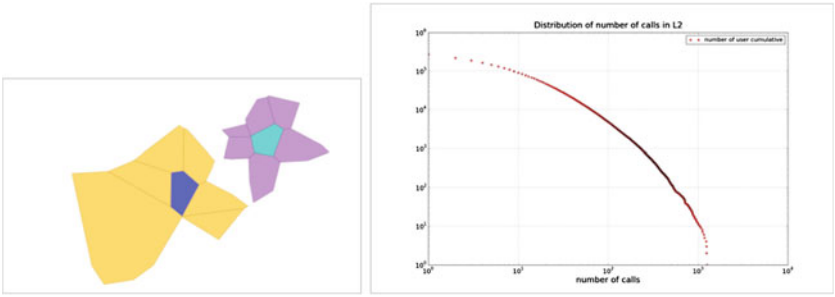

Fig. 2. Example of L1 and L2 detection (left), distribution of number of calls in the second most frequent location (right)

The L2 is defined as the area that ranked second in terms of call frequency, excluding all areas already absorbed into L1. Figure 2 Shows a visual example of L1 and L2. While in the example L1 and L2 are quite distant (for instance, these individuals might come to the city to work), in some cases they might be adjacent, which happens quite often in city centers where antennas are more dense. Also in this case it is necessary to consider a minimum support of phone calls to identify the significance of the places identified based on the distribution shown in Fig. 2(right). The result of this analysis will be used in the next step for the study of the systematic movements between preferred locations.

4.2 Detecting Systematic Movements

The focus of this analysis is the detection of systematic movements considering two separated time frames: a morning time frame, and an afternoon time frame, in which the users usually move, respectively, from home to work and from work to home. The first step is to identify the movements performed by individuals from L1 to L2 (L1 → L2) and from L2 to L1 (L2 → L1). It is important to notice that we are looking for movement between these two areas even if they are not contiguous, i.e. other areas were traversed between them, as shown in Fig. 3 (A is distinct from L1 and L2).

The second step consists in selecting only the systematic movements, which is done by applying two different constraints: (i) request a minimum number of movements between the pair; and (ii) request a minimum value for the *lift* measure [11] of the pattern L1 → L2, which we define as:

$$LIFT(L1, L2) = \frac{P(L1 \wedge L2)}{P(L1) \cdot P(L2)} \tag{1}$$

where $P(Li)$ $(i \in \{1, 2\})$ represents the frequency of Li, expressed as fraction of days where it appears at least once, and $P(L1 \wedge L2)$ represents the frequency of L1 and L2 appearing together. Lift measures the correlation between L1 and L2, resulting high if they appear together often w.r.t. the frequency of L1 and L2

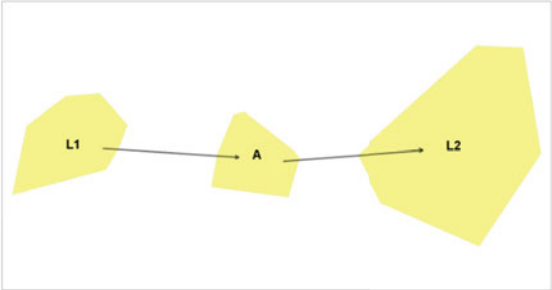

Fig. 3. Example of flow from L1 to L2

taken separately. The main purpose is to normalize the frequency of L1 → L2 w.r.t. the frequency of calls of the user, since otherwise the candidate movements of frequent callers would be excessively favoured in the selection. The threshold on the number of movements is usually adopted in literature to exclude extreme cases where the lift (or other correlation or relative frequency measures) is not significant. More important is the LIFT measure, in fact to select. In our case, after a preliminary exploration we chose to select only pairs that appeared at least 3 times. The threshold for the lift measure was chosen based on the study of its distribution, selecting the value where the slope of the cumulative distribution begins a sudden drop, corresponding to 0.7.

4.3 Systematic OD Matrix

As previously mentioned, the final goal of our analysis is the synthesis of O/D matrices that summarize the expected traffic flows between spatial regions. Our O/D matrices will focus systematic mobility, which represents the core (though not the only) part of traffic. In Fig. 4 some examples of intra-city traffic are shown, the first one from one origin to several different destinations, the second one from several origins to a single destination.

5 Application to the Case Study

In this section we summarize the process and results of inferring a transport model for Ivory Coast by applying the systematic mobility demand model extracted through the methodology illustrated in the previous section.

Network. We used data of the OpenStreetMap (OSM) road networks for Ivory Coast and Abidjan as a base for modeling. Although of great detail,in these network data many links are not, or not properly, connected. For route calculation this evidently is problematic. We therefore used an algorithm in a Geographic Information System (GIS) to find unconnected or badly connected roads and connected them properly or at least logically. Furthermore

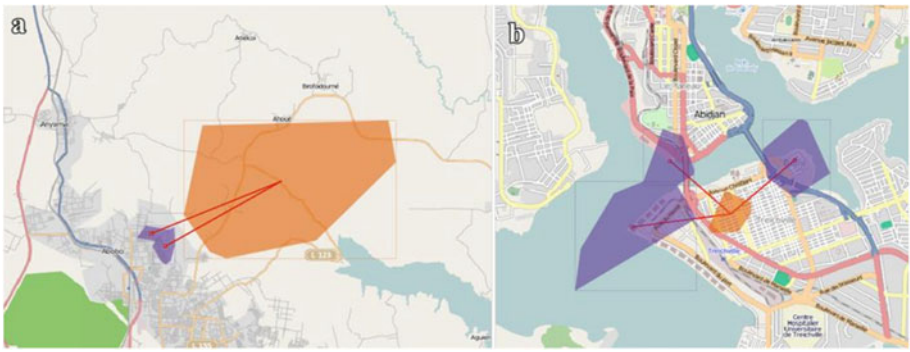

Fig. 4. Examples of traffic flows taken from one of the O/D matrices produced: (a) flows from the outskirts of Abidjian to the city; (b) from a single area to other districts.

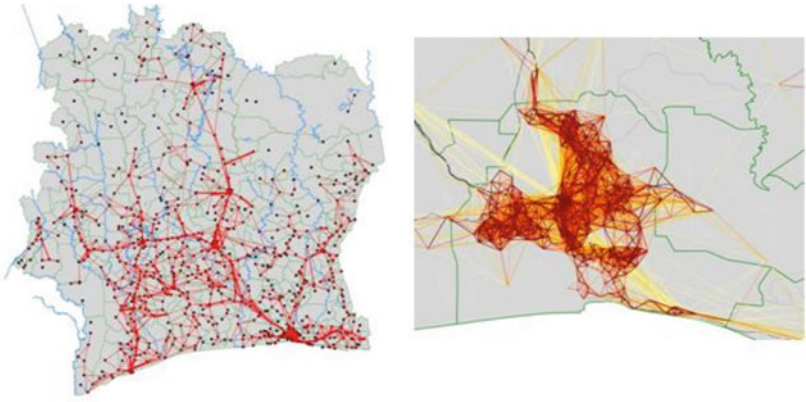

Fig. 5. Mobile phone movements in Ivory Coast and Abidjan.

we programmed an algorithm to identify small island-networks. They were either connected to the main network, typically we added a few ferries to connect real islands, or erased because they seemed illogical or unimportant. The resulting digital road network was imported in OmniTRANS, the software for transport modeling. Network link attributes, such as speed, which are necessary to calculate the shortest route between an origin and destination, were derived from link type information which is available from the OSM-network. The next step was to connect the antennas, e.g. the data carriers, to the network. In anticipation of this step we did not remove all the small roads from the network, which is often done in transport models. By keeping them we could simply connect each antenna to the nearest road. Of course this gives an overload on that particular minor road, but this quickly flattens out as all trips divert in different directions and converge on the major roads.

Fig. 6. Traffic model for 24 h period for Ivory Coast (left) and Abidjan area (right)

OD Matrix. The previously described OD-matrices derived from mobile phone data were imported in the transport model using a simple format (origin, destination and number of trips between them). Different OD-matrices were established for different time periods: an AM peak period (5–11), a PM peak period (15–21) and a 24h-period. Without assignment to a network these data already provide interesting images of movements. Figure 5 shows tower locations in Ivory Coast and movements of mobile phones between tower locations for Ivory Coast and the Abidjan area. These figures already provide a rough idea of the road network and major flows. However, a transport model is needed to gain insight into flows on the road network.

Assignment. We used OmniTRANS V6 software to assign OD-matrices to the road network. A simple all-or-nothing assignment technique was used by which all trips on an OD-relation are assigned to the calculated shortest route in time between the origin and the destination. More sophisticated assignment techniques are available (though requiring more data, like the road capacity), which take into account that different routes may be chosen because of congestion, but we felt the use of more sophisticated assignment technique was beyond the scope of this research.

Results and Interpretation of the Transport Model. Figure 6 shows assignments of the OD-matrices based on the GSM data for a typical 24 hour period for Ivory Coast and the Abidjan area. All major, national and urban transportation corridors are immediately visible from these plots. Also, the comparison between the linear movements between cell towers in Fig. 5 and the assigned movements in Fig. 6 provides a clear impression of the added value of assigning the movements to the road network. For purposes of readability we have decreased the level of detail in the figures in this paper. The original model plots allow much more detailed analysis of volumes on road links, in both directions of roads. Figure 7 shows traffic assignment on the network for the morning peak period in the greater (left) and central Abidjan area (right). As can be seen from all assignments the absolute flows, or traffic volumes on the network are very low and do not represent real traffic volumes on the network, since they are based on a selection of the

sample provided for this study. Still, as a relative measure all figures show roads with more and less dense traffic. To make the transport model suitable for identifying and exploring current or future transport problems in Ivory Coast, an accurate assessment of absolute traffic flows on relevant parts of the network is necessary. In the following discussion we will deal with what must be done to overcome the limitations of the current model.

Discussion. It is highly rewarding that we were able to create a transport model for an area where we, as researchers, have never been, and for which data were solely obtained from the internet and from a completely new source (GSM call traces). However, the transport model is not yet finished, and this is mainly due to a number of remaining limitations in the data which were available. The good news is, that all these limitations can be solved, but the effort to do this varies.

First, to make an estimate of actual and validated traffic flows on the network, the current model values should be augmented or weighed by a factor. This factor will depend for instance on the market penetration of Orange operated cellphones, cellphone ownership and cellphone usage in Ivory Coast etc. As these levels may differ throughout the country, the weighing-method should take into account local differences. Techniques to do this are already available since also in transport modeling using conventional data, OD matrices are calibrated from traffic counts. A set of reliable traffic counts in the network should therefore suffice to establish augmented OD matrices to describe not just the traffic based on mobile phones traces, but the total traffic flows.

A second, more serious limitation in the present model is that it does not make a distinction between the different transport modalities. To achieve such a distinction filters and algorithms must be developed to detect and estimate different modes of travel from data. The data available now would hardly allow a distinction of modes that is based on travel speeds. However, for modeling purposes, some basic statistics on modal split can be obtained from the same traffic counts as are needed to augment the traffic present in the model and could highly increase the quality of the model.

If we could overcome the above mentioned problems, and it seems absolutely feasible that this can be done, then we will have a model that can be used both for an accurate assessment of the current traffic situation in CI and Abidjan and for forecasting purposes. For forecasting purposes one needs to implement future planned projects in the model and specify the expected general growth in mobility for all modes. Once we have an adequately augmented mode specific OD matrix, the model immediately allows other calculations and forecasts for environmental impact analysis, such as greenhouse gasses, NOx and PM10 emissions. Of course, for these purposes additional statistics on the Ivory Coasts vehicle park must be incorporated.

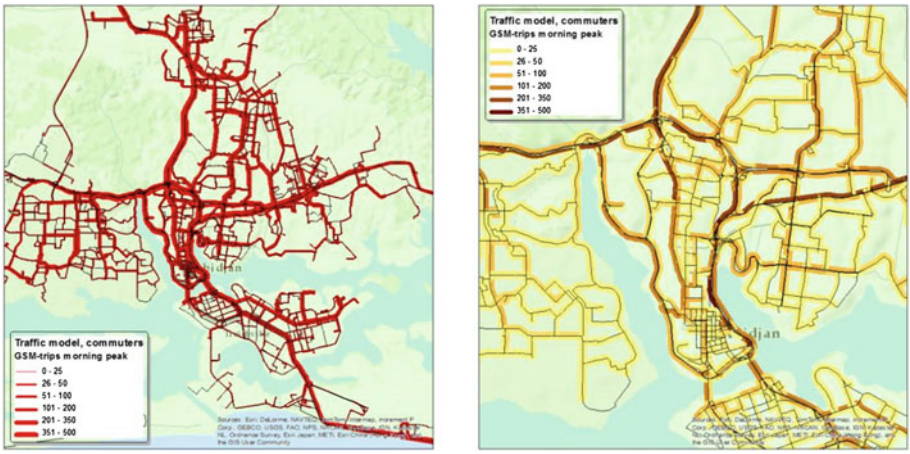

Fig. 7. Traffic model for morning peak period for greater Abidjan (left) and Abidjan Centre Area (right).

6 Conclusions

The present study shows that even with limited data from GSM traces, as in the case study considered, it is possible to derive valid information on systematic mobility behavior of people between frequently visited locations for areas that lack information on mobility. From these mobile phone data, origin-destination tables can be created for a chosen geographical area, which can then be used as an input for a transport model of the area. The fact that this can be done, overcomes one of the serious hurdles that until now have impeded the use of transport models in many developmental countries: lack of data. It is therefore absolutely worthwhile to take this proof of concept one step further, and create and validate a transport model that can indeed be used by public authorities, engineering firms, investors etc. in developmental countries. In this paper we discussed the most important steps which need to be taken.

References

1. Olteanu, A.-M., Trasarti, R., Couronn, T., Giannotti, F., Nanni, M., Smoreda, Z., Ziemlicki, C.: GSM data analysis for tourism application. In: ISSDQ (2011)
2. Trasarti, R., Pinelli, F., Nanni, M., Giannotti, F.: Mining mobility user profiles for car pooling. In: ACM KDD (2011)
3. Csji, B.C., et al.: Exploring the mobility of mobile phone users. Report on arXiv: 1211.6014 [physics.soc-ph]
4. Giannotti, F., et al.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. (Special issue on Data Management for Mobile Services) **20**, 695–719 (2011)
5. Csjia, B.C., et al.: Exploring the mobility of mobile phone users. Physica A : Stat. Mech. Appl. J. **392**(6), 1459–1473 (2013)

6. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Identifying users profiles from mobile calls habits. In: UrbComp (2012)
7. Hensher, D.A., Kenneth, J.: Handbook of Transport Modelling. Pergamon, Amsterdam (2000)
8. Ortúzar, J., Willumsen, L.G.: Modelling Transport, 4th edn. Wiley, West Sussex (2011)
9. Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating origin-destination flows using mobile phone data. Pervasive Comput. **4**, 36–44 (2011)
10. D4D Challenge. http://www.d4d.orange.com/
11. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Boston (2006). ISBN 0-321-32136-7

http://www.springer.com/978-3-319-04177-3

Citizen in Sensor Networks
Second International Workshop, CitiSens 2013, Barcelona, Spain, September 19, 2013, Revised Selected Papers
Nin, J.; Villatoro, D. (Eds.)
2014, IX, 109 p. 33 illus., Softcover
ISBN: 978-3-319-04177-3