Kujala, Rainer; Aledavood, Talayeh; Saramäki, Jari

Estimation and monitoring of city-to-city travel times using call detail records

Published in:
EPJ Data Science

DOI:
10.1140/epjds/s13688-016-0067-3

Published: 01/12/2016

Document Version
Publisher's PDF, also known as Version of record

Kujala et al. *EPJ Data Science* (2016) 5:6
DOI 10.1140/epjds/s13688-016-0067-3

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Estimation and monitoring of city-to-city travel times using call detail records

Rainer Kujala*, Talayeh Aledavood and Jari Saramäki

*Correspondence:
Rainer.Kujala@aalto.fi
Department of Computer Science, Aalto University, P.O. Box 15400, Espoo, FI-00076, Finland

Abstract

Whenever someone makes or receives a call on a mobile telephone, a Call Detail Record (CDR) is automatically generated by the operator for billing purposes. CDRs have a wide range of applications beyond billing, from social science to data-driven development. Recently, CDRs have been increasingly used to study human mobility, whose understanding is crucial *e.g.* for planning efficient transportation infrastructure. A major difficulty in analyzing human mobility using CDR data is that the location of a cell phone user is not recorded continuously but typically only when a call is initiated or a text message is sent. In this paper we address this problem, and develop a method for estimating travel times between cities based on CDRs that relies not on individual trajectories of people, but their collective statistical properties. We apply our method to data from Senegal, released by Sonatel and Orange for the 2014 Data for Development Challenge. We turn CDR mobility traces to estimates on travel times between Senegalese cities, filling an existing gap in knowledge. Moreover, the proposed method is shown to be highly valuable for monitoring travel conditions and their changes in near real-time, as demonstrated by measuring the decrease in travel times due to the opening of the Dakar-Diamniadio highway. Overall, our results indicate that it is possible to extract reliable *de facto* information on typical travel times that is useful for a variety of audiences ranging from casual travelers to transport infrastructure planners.

Keywords: data for development; call detail records; mobile phones; travel time estimation; near real-time monitoring

1 Introduction

Mobile phones are ubiquitous, widely available and used all over the world. They have also proven to be an invaluable source of high-quality data for studying different aspects of human societies [1–3], especially for development purposes [4, 5]. Such studies typically use Call Detail Record (CDR) data that are collected by telecommunication operators for billing purposes and therefore come with no extra cost or overhead. CDRs contain information on communication events such as calls or text messages, including the initiator and recipient, time of contact, and which cell tower is involved in the contact.

Studying CDRs has been especially helpful for developing and underdeveloped countries, where there is often a lack of systematic population-level data collection, or in the aftermath of natural disasters, where individuals are hard to reach or their location is unknown [6]. In the recent years, global entities like UN Global Pulse have published re-

ports on use of these data for such purposes [4], and telecommunication companies such as Orange and Telecom Italia have set up data challenges for scientists to study CDRs for development purposes [7, 8].

One important line of research applies CDR analysis to study human mobility and transportation and to develop methods that can be used *e.g.* for urban real-time monitoring [9] and planning [10], or for optimization of transportation infrastructure [11–14].

The time it takes to travel between different locations is a key constraint (and a key descriptor) of human mobility. Thus, up-to-date information on *de facto* travel times is not only of importance to travelers, but also for planning and governance of transport infrastructure. For instance, such information could be useful for monitoring road conditions or assessing access times to hospitals. The importance of *de facto* travel times is therefore evident, but their availability is still limited, in particular in developing countries, due to a lack of available resources required for such monitoring.

In practice, there are many ways to estimate and monitor travel times. Travel time information can be estimated with different techniques typically used by transportation engineers, ranging from magnetic loop detectors, automatic register plate recognition systems, and recording of GPS traces to traditional surveying methods [15–17]. Although some of these methods provide highly accurate real-time estimates on travel speeds and times, they typically require installation of physical equipment (*e.g.* magnetic loop detectors) which makes them resource-intensive, or they are labor-intensive (surveys). GPS-based methods require less resources. In particular, Google or other vendors of smartphone operating systems can easily leverage on an existing population of suitable devices to collect raw data for computing travel time estimates. However, even though mobile phones are common in developing countries, smartphone penetration is typically low [18], making the collection of data from GPS-enabled smartphones difficult in practice. Also there may be no commercial interest in providing detailed, high-quality information on travel times in developing countries. Furthermore, algorithms used for extracting travel time estimates from raw data are not typically available.

One alternative approach to estimating travel times is to use data generated by communication between a mobile phone and the cellular network base stations. In developed countries, the most important use case is to provide accurate real-time information on traffic conditions and therefore most studies and commercial projects have been focused on this topic [19–22]. There are two main approaches for estimating travel time information using mobile phones and the cellular network. The first uses information generated when mobile phones move across the coverage areas of cell-towers, which results in handovers and location area change events [23–25]. The second is based on signaling strengths and delays between a mobile phone and nearby cell towers [26, 27]. When done periodically, this results in GPS-like coordinate trajectories which can then be further refined into travel time distributions between origin-destination (OD) pairs [27].

While many of these systems have also been commercially implemented [19, 28], such travel time estimation systems are not yet adopted worldwide and even in some developed countries they are still at the pilot phase [29].

To summarize, in the context of travel time estimation, most studies focus on providing accurate, real-time estimates on specific road segments while less attention has been given to the travel times actually experienced by the users on longer trips. Additionally, most methods are either costly, labor intensive or rely on infrastructures which are non-existent

in developing countries. Therefore, there is a need for methods which (1) are inexpensive and are not resource or labor-intensive (2) do not depend on complicated infrastructure or hardware (3) provide accurate estimates of travel times experienced by users.

In this paper, we show that this can be achieved with the help of CDR data already stored for billing purposes, without the need for implementing more detailed hand-over or triangulation data analysis pipelines. The benefit of using billing data is that mobile operators always collect this data in a standardized format; there is even a dedicated software package for analyzing such data (http://bandicoot.mit.edu/) [30]. However, extracting accurate travel time information from CDRs is not a straightforward task because the location of a user is recorded only when the user initiates a call or sends a text message. Therefore, a single CDR-based mobility trajectory is typically very sparse in time, and cannot directly be used for estimating travel times between locations. However, when multiple mobility trajectories are pooled and analyzed as a whole, it turns out it is possible to produce reliable travel time estimates.

To this end, we have developed a method for automated extraction of typical travel times between cities from CDR data. Due to the simplicity and low computational cost of our method, we are immediately able to scale it up to the country level instead of the more local scales typical for other methods. The method aims at providing an overall view on travel times between cities and it enables monitoring of travel times and conditions in the long term. It has been especially designed for developing countries where reliable information on travel times and transport infrastructure is limited or not available at all. Unlike some of the above-mentioned methods, it is not designed for producing real-time traffic speed estimates for specific road segments; however, as we show with an example piece of Senegalese highway, it does allow detecting sudden changes in travel times.

The data we analyze originates from Senegal, for which Orange and Sonatel have provided anonymized CDR-based mobility data sets in conjunction with the 'Data for Development Challenge 2014' (D4D Challenge) [8]. To show our method's performance in practice, we compare our results to existing travel time information available from alternative sources, such as the travel times provided by Google. Furthermore, to demonstrate that the method is capable of monitoring changes in travel condition in near-real-time, we estimate how much the opening of the Dakar-Diamniadio highway dropped the typical travel times between the capital Dakar and the nearby city of Pout.

2 Data and methods

2.1 Data

In this study, we have used 25 anonymized mobility data sets provided by Orange and Sonatel for the 2014 D4D Challenge [8]. Each set contains ∼300,000 mobility traces for a two-week time span; a mobility trace contains the cell tower IDs and time stamps of calls and text messages made by one anonymized customer. In the provided data set, users whose traces span less than 75% of the days of in a given two-week period have been filtered out, together with users who have more than 1,000 weekly events and likely correspond to non-human users such as machines sending text messages. Both filtering processes have been performed by the D4D Challenge organizers, *i.e.* at data source. As shown in Figure 1(A), after the filtering most users have been observed between 10 and 1,000 times during a two-week time span. Because of privacy and commercial reasons, only approximate coordinates are given for the locations of the cell towers, and the time

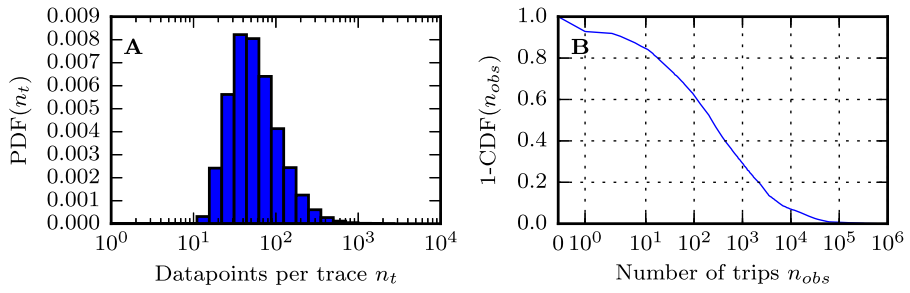

Figure 1 Mobility trace lengths n_t and inter-observation time counts between origin-destination city pairs. In Panel **A**, we show the logarithmically binned distribution of the number of data points in the users' CDR-based mobility traces. The distribution shows that most two-week mobility trajectories have between 10^1 and 10^3 data points. In Panel **B**, we show the complementary cumulative distribution (1-CDF) of the number of inter-observation counts for all origin-destination city pairs. Note that a large number of city pairs (∼10%) have zero inter-observation times, which is partially due to some cities not being allocated to any (constantly) active cell tower.

resolution of data is restricted to 10 minutes. For a more detailed description of the data set, see Ref. [8].

No data are perfect, and this data set is no exception. The biggest problem arises from the fact that the data only contains the locations of the cell towers at the end of the data collection period.[a] Thus, changes in the cell tower or, to be precise, in the locations of the base transceiver stations during the year can go unnoticed, and cause errors in the data. To reduce errors caused by cell towers (or their IDs) whose location has changed and by cell towers that were introduced during the time span of the data set, we have only included in our analysis those cell towers that were associated with at least one CDR entry on each day of the time span covered by the data. This white-list of cell towers was created using another data set that contained the hourly numbers of calls and text messages sent and received at each cell tower. In total the white list contained 1,093 cell towers out of the total number of 1,666 provided in the data set. Although this tower-level filtering of the data has helped to reduce errors, it is still apparent in some of the results that the source data comes with erroneous tower locations.

Possible problems with tower locations are also mitigated by focusing on cities instead of individual cell towers. To this end, we obtained a list of 62 major Senegalese cities and their geo-coordinates from www.tageo.com [31]. With the help of this information and the provided cell tower locations, we assigned a set of cell towers to each city, such that each cell tower is assigned to its closest city whenever their distance is at most 10 km. Note that two cities (Wassadou and Ourossogui) out of the total of 62 cities were not assigned any cell towers. The locations of the cities and their associated cell towers are displayed in Figure 2(A).

2.2 Determination of typical travel times

Given two cities i and j and their corresponding sets of cell towers I and J, we say that user u has made a *trip* from city i to city j whenever the mobility trajectory of u first contains one of the cell towers in I, and at a later point one of the cell towers in J. Between the start and end of a trip, a user can visit any other cities and cell towers, but can not visit any cell towers corresponding to the origin or the destination city. Thus a trip from i to j consists of a series of time-ordered observations (time, user, tower ID), where the first

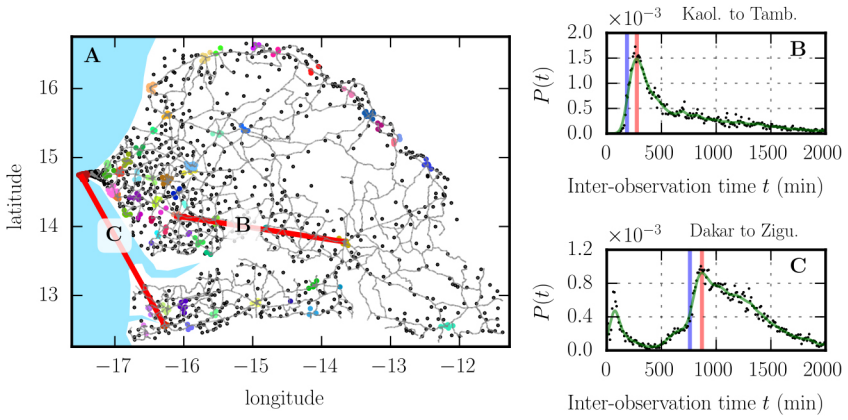

Figure 2 Extraction of typical travel times between Senegalese cities. Panel **A**: Each set of cell towers assigned to a city are colored with same color. Black dots indicate cell towers that were not assigned to any city. Light blue background indicates sea, and white background land. Note also that southern and northern parts of Senegal are separated by Gambia, for which road data is not shown. Panels **B** and **C**: Two examples of inter-observation time distributions (B: from Kaolack to Tambacounda; C: from Dakar to Ziguinchor). The empirical inter-observation time distributions are shown with black dots, and the green curve represents our kernel density estimate of the inter-observation time probability density. The red vertical lines indicate the estimated typical travel time corresponding to the peak, and the blue vertical lines indicate the lower bound estimates. In Panel B, we see a typical pattern with a single clear peak that is located at 275 minutes (4 h 35 min). This gives us an estimate of the typical travel time from Kaolack to Tambacounda. In Panel C, however, there are two peaks. The first peak is located at ∼100 minutes and is presumably from air traffic between Dakar and Ziguinchor. The second peak is located at 870 minutes (= 14 h 30 min), which matches well with the travel time taken to reach Ziguinchor from Dakar with ferry (15 h).

tower ID belongs to the set I and the last tower ID to the set J. The *inter-observation time* corresponding to each trip is then defined simply as the time between the first and last observation. Note that a user can be simultaneously on multiple trips. A schematic example of a city-level trajectory, and the resulting inter-observation times are presented in Figure 3. In Figure 1(B) we show the pooled distribution of the number of extracted inter-observation times for the 62×61 OD-pairs.

To estimate the typical travel time from city i to city j, we pool all inter-observation times from the mobility trajectories of different users, and investigate their distribution. In theory, the shortest observed inter-observation time would be indicative of how fast one can travel between the two cities. However, the shortest inter-observation time may not represent the *typical* travel time between these two locations and it is also particularly sensitive to any errors in the data. Because of this, we focus on the peak of the inter-observation time distribution instead.

For accurately estimating the location of the peak, some smoothing of the inter-observation distribution is necessary as the original distribution of data can fluctuate a lot, even though our data is already binned to 10 minute intervals due to the restricted temporal resolution of 10 minutes. Smoothing is of most importance when an OD-pair has a low number of inter-observation times (see Figure 4 for an example).

It is also worth noting that travel can take place using different modes of transportation which can cause multiple peaks in the inter-observation time distribution. This can clearly be seen in Figure 2(C), where there is first a peak corresponding to travel by air followed by a peak corresponding to travel by sea and land. In this study we focus on the most typical travel modes. As travel by air is not very common within Senegal [32], we focus

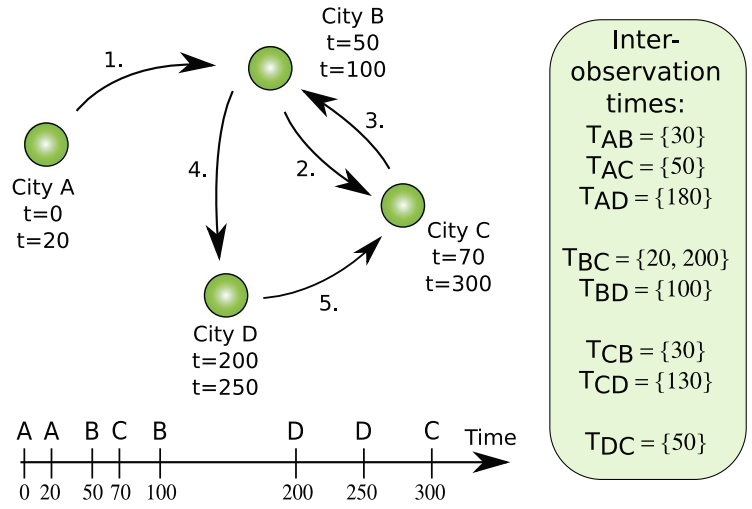

Figure 3 A schematic representation of one user's mobility trajectory and the resulting inter-observation times. On the left, the mobility trajectory of one person is visualized both as a spatial representation (top) and as a timeline (bottom). In the spatial representation each circle corresponds to a city, and an arrow corresponds to a movement from one city to another. The ordering of the movements is indicated by ordinal numbers and the times when the user has been observed in each city are shown below the name of each city. The timeline presentation below shows the same information in a more compact form. On the right, we have also listed all inter-observation times that can be computed from the trajectory.

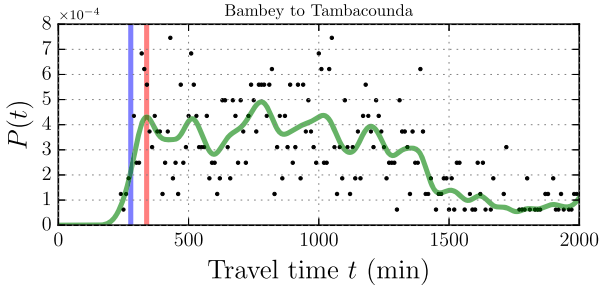

Figure 4 Importance of smoothing. In the figure we show an example OD pair with a low number of extracted inter-observation times. The black dots show the original distribution of data consisting of in total 1,609 inter-observation times. The green curve shows the smoothed estimate of the probability density function, while the red vertical line denotes the peak estimate and the blue vertical line denotes the lower bound estimate. As is evident from the figure, the original data fluctuates a lot and the general trend of the data is better visible in the smoothed distribution. Furthermore, our decision rule for the peak now selects the first peak of sufficient magnitude, which is a more reasonable estimate than the even higher peak located around 750 minutes.

on travel times corresponding to straight-line travel speeds of less than 100 km/h which should allow for all different travel modes by sea and land but filter out air traffic. Note that given the typical road and travel conditions and that the limit is on straight-line speeds, this manually chosen limit is rather generous and will almost certainly not exclude any actual land travel trips. This thresholding also helps to further filter out some erroneous results due to irregularities in the source data.

Because of various biases in the data (*e.g.* different mobile phone usage frequencies leading to different waiting times before first call is made at destination, or tower location offsets) there is no guarantee that the location of a peak in an inter-observation time dis-

tribution would precisely correspond to a typical travel time between two cities. Thus, information on the peak's width is also important. The right-hand side of the peak typically decays slowly as there is no natural limit to a trip's duration. Therefore we focus on the left-hand side of the peak and its *lower bound*. If the position of the peak is considered as an estimate of the typical travel time, the lower bound measures the best case, travel under optimal conditions. The lower bound is computed as follows: given a peak's location t_p, the location of its lower bound t_l is defined as the largest inter-observation time such that (i) $t_l < t_p$, and (ii) the value of the smoothed inter-observation time distribution is lower than or equal to half the peak height.

In detail, our analysis pipeline for estimating typical travel times between cities is as follows:

1. *Compute inter-observation time distributions*

   Loop through the CDR data, compute inter-observation times for each origin-destination city pair, and pool the results into inter-observation distributions.
2. *Smooth the distributions*

   To smooth the inter-observation time distributions, use a standard Gaussian kernel

   $$G(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} \tag{1}$$

   with a standard deviation σ corresponding to 30 minutes. The bandwidth of 30 minutes was chosen as it was found to allow reasonable travel time estimations for city pairs with fewer trips, while not oversmoothing the original data. The smoothed density estimates $P_s(t)$ can then obtained from the original inter-observation time distribution $P_o(t)$:

   $$P_s(t) = C \sum_{t'=0}^{t_{\max}} G\big(t - t'\big) P_o(t'), \tag{2}$$

   where $t_{\max}$ is the largest inter-observation time permitted by the data (2 weeks) and C is a normalization coefficient guaranteeing that the final smoothed distribution $P_s(t)$ is a valid probability density function. The smoothed density estimates are evaluated at 1 min intervals.
3. *Find all maxima*

   Find all local maxima of the smoothed probability density functions whose corresponding straight-line travel speed does not exceed 100 km/h (to filter out air traffic and errors in the original data). This can be done simply by going through the elements of the vector of smoothed density estimates: an element is a local maximum when its value is higher than those of its neighbors.
4. *Detect the peak corresponding to typical travel time*

   From each smoothed probability density functions, select the peak with smallest travel time such that the height of the peak is at least 0.5 times the height of the largest peak fulfilling the travel speed restriction. Typically this condition results in simply choosing the highest peak of the distribution, but with origin-destination city pairs with a low number of observations this condition was found to provide more robust results (see Figure 4).

5. *Compute the lower bound estimate*

   Select the closest point in time smaller than the peak time such that the height of the distribution at this point is smaller than or equal to half the height of the detected peak. In case the inter-observation time distribution does not fall to half peak height on peak's left hand side, set the value for the lower bound to 0 minutes. Typically, such cases are due to irregularities in the source data.

Note that because there is no ground-truth calibration data, we are forced to set some parameters of the method on the basis of reasonable assumptions, instead of adjusting their values based on calibration. In the following sections, we will nevertheless discuss possible causes of estimation biases. Especially, we will investigate the biases caused by varying the smoothing bandwidth.

The code implementing the above analysis pipeline for extracting typical travel times from CDR data is freely available at https://github.com/rmkujala/d4dttimes.

2.3 Estimation biases

Our peak and lower-bound estimates are, of course, prone to different kinds of biases due to our definition of inter-observation times. First, each inter-observation time typically includes not only the actual time of travel but also period before and after, as calls are not made exactly on departure or arrival. This bias is however difficult to correct, as it is not known how mobile phone usage and travel behavior are coupled (and inter-call times are typically very broadly distributed, see *e.g.* [3]). Moreover, we do not filter out detours taken by travelers which cause the long tails in the inter-observation times as seen in Figure 2. On the other hand, the range of cell towers can cover areas that are far from the location of a city, which can shorten inter-observation times between pairs of cities. These individual biases can thus sum up to a bias that can be either negative or positive, and that is difficult to estimate using CDR data alone. However, if calibration data *e.g.* based on GPS recordings of individuals were available, it should be possible to correct for these different biases. Nevertheless, our example cases will show that our estimates tend to be close to quoted travel times found from literature. In any case, it seems reasonable to assume that the bias remains relatively constant for any pair of cities, and thus when the method is used for monitoring changes in travel times, possible biases no longer matter.

2.4 Effect of the number of samples on the estimates

As our method relies on the distribution of inter-observation times between two cities, it is important to know how much data is required for reliable estimates. To get some idea of the amount of data required for robust travel time estimation, we investigated how the estimation error decreases with the number of data points. This was done by bootstrap resampling the original inter-observation time distributions so that bootstrap sample sizes ranged from 100 up to the total number of data points in the sample. For each sample size, we calculated 1,000 bootstrap estimates and computed the median as well as the 5th and 95th percentiles of the bootstrap estimate distributions. Here, we report the results obtained for the two city pairs presented in Figure 2 ('Kaolack to Tambacounda' and 'Dakar to Ziquinchor'). In addition, we also investigated two origin-destination city pairs ('Dakar to Thies' and 'Dakar to Kaolack') for which a very large number of inter-observation times ($>10^5$) were available when data were aggregated over the entire data collection period.

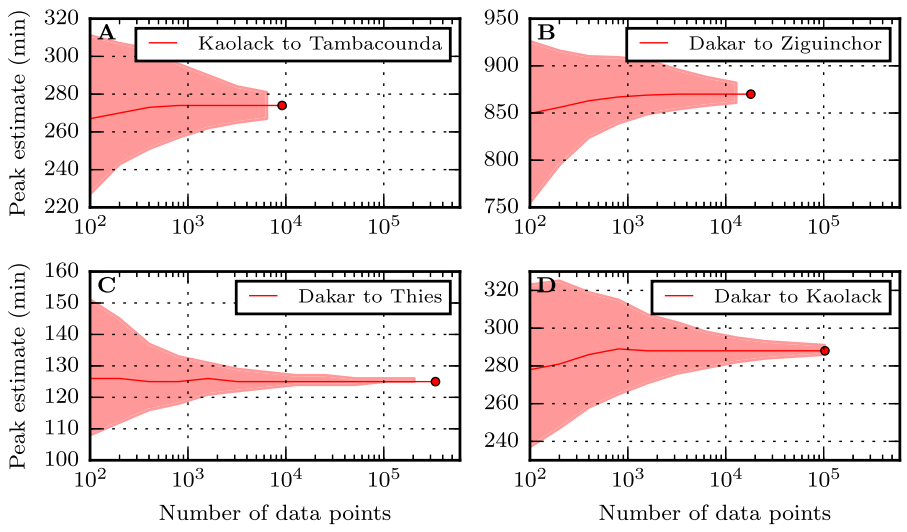

Figure 5 Dependence of estimation accuracy and bias on the number of data points. Solid lines denote median estimates, and the shadowed area denotes the 5th and 95th percentiles of the bootstrap estimate distributions. The circle at the end of each curve denotes the travel time estimate obtained using the full inter-observation time distribution.

The results are shown in Figure 5, and they illustrate two main points: First, our estimates on the location of the peak are relatively unbiased when at least 1,000 inter-observation data points are available, as the median of the bootstrap estimates remains close to the final value of the full distribution after this limit. Second, based on the 5th and 95th percentiles of the bootstrap distributions, to reach an acceptable 5 min estimation accuracy we need of the order of 10,000 data points. In our data set, we find 266 origin-destination city pairs (6.9% of all origin-destination pairs) that fulfill this criteria when analyzed over the whole year. For the full distribution on the number of inter-observation data points, see Figure 1(B).

Naturally, as the shape of inter-observation time distributions differs across city pairs, the amount of data required for accurate travel-time estimates may vary. As a rough rule of thumb, we nevertheless conclude that approximately 10,000 inter-observation times are required for each origin-destination pair for obtaining reliable estimates. It is also worth stressing that this rule of thumb is specific to this study only, as the mobility data used here came in two-week chunks limiting the longest possible observable inter-observation times accordingly.

2.5 Effect of the width of the smoothing kernel

Next, we discuss how the width of the Gaussian kernel used for smoothing the inter-observation time distributions affects the results. In this work we report results with a Gaussian kernel of width that corresponds to 30 min in standard deviation, which we found to yield reasonable results. In general, it would be good to select kernel width adaptively *e.g.* using cross validation. However, given that the time resolution (10 min) of the data was artificially heavily limited, this would have not been very straightforward.

We have nevertheless investigated how the smoothing bandwidth used affects our results. To this end, we computed the peak and lower bound estimates with a range of different smoothing bandwidths for origin-destination city pairs that had at least 10,000 inter-

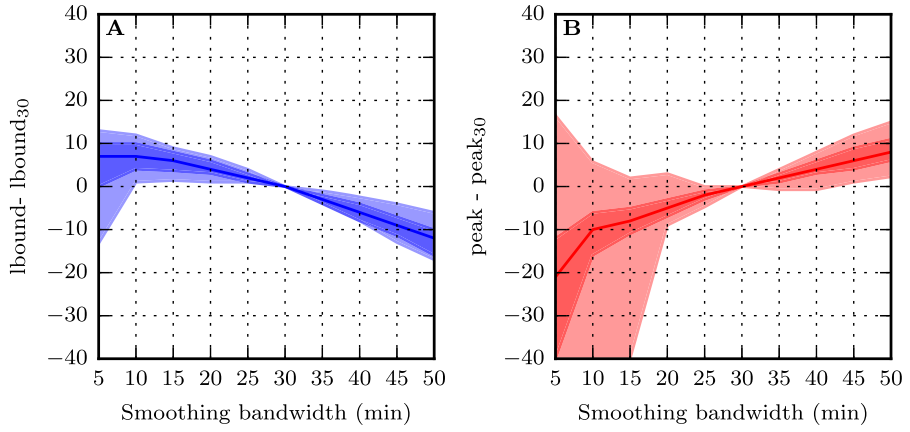

Figure 6 The effect of the smoothing bandwidth on the travel time estimates. In Panel **A** we show the distribution of the lower bound estimates normalized by subtracting the lower bound estimate obtained with smoothing bandwidth of 30 minutes. In Panel **B** we show the distribution of similarly normalized peak estimates. In both panels the outer shaded area denotes the 5th and 95th percentile, the inner shaded area denotes the 25th and 95th percentiles of all estimates, and the solid lines correspond to the median of all the normalized estimates. In these plots, we only show data for OD pairs with at least 10,000 data points. Furthermore, the results for the lower bound estimates are only based on those results for which we are able to identify the lower bound with all different bandwidth values. (With some OD pairs the inter-observation time distribution never falls to half of the identified peak's width on the peak's left hand side due to data irregularities.) Panel B shows that most peak estimates tend to stabilize when the width of the smoothing kernel reaches 20 min, as is shown by the 5th percentile of the normalized estimates. Additionally, we note that smoothing causes a systematic bias to the results: the larger the kernel width, the larger are the peak estimates and the smaller are the lower bound estimates.

observation times within the time span of the data. We arrived at two main conclusions (see Figure 6): First, our estimates seem to stabilize when the width of the smoothing kernel reaches 20 min (this becomes more emphasized when the threshold is set to 1,000 inter-observation times). Second, our lower bound estimates tend to decrease and peak estimates increase when the smoothing bandwidth increases. This is due to the skewness of the inter-observation time distributions: As the right tail of the distribution is typically fatter than the left tail, smoothing systematically shifts the peak to the right and the lower-bound estimate to the left.

3 Results

3.1 City-to-city travel times within Senegal

We begin by reporting results on travel time estimates extracted for all origin-destination pairs for which at least 10,000 inter-observation times were discovered. In the supplementary web-page (see Additional file 1), we also report our estimates and bootstrap error bounds for all origin-destination city pairs for which at least 1,000 inter-observation times were available.

As any official information on times of travel between Senegalese cities is scarce and hard to find, there is no obvious ground truth available for validating our results. Nevertheless, to give our estimates general credibility, we performed two different sanity checks. First, the travel time estimates should be symmetric: the estimated travel time from city i to j should be approximately equal to the travel time from city j to i. As shown in Figure 7(A), our estimates do generally fulfill this condition. Second, if we take up a simplistic assumption of constant average travel speed throughout the country, we would expect an

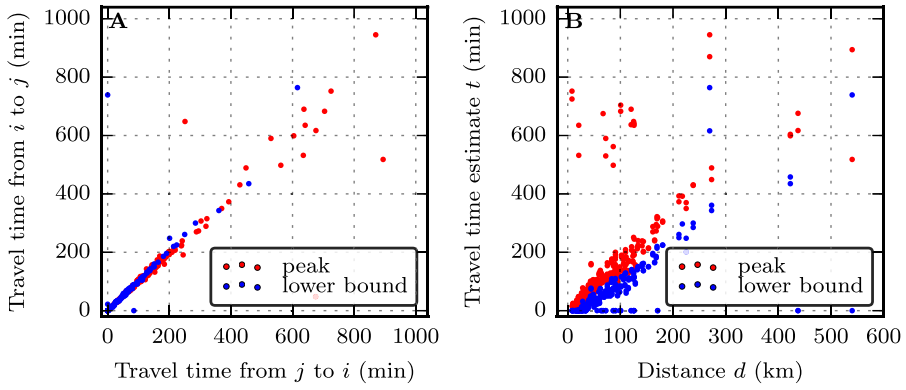

Figure 7 Symmetricity and scaling of travel time estimates with distance. In Panel **A**, we present a scatter plot of estimated travel times from *i* to *j* versus times from *j* to *i*. It is seen that the estimates are rather symmetric, as they should. Note that each pair of cities corresponds to a single point in the plot, and the selection of which of the cities corresponds to *i* and which to *j* is made arbitrarily. In Panel **B**, we show how the estimates scale with distance. Overall, we can observe an approximately linear trend. There are a few points that correspond to clearly infeasible estimates due to erroneous data that seem to have very long travel times (500-800 minutes) even though the distance between city pairs is low (0-150 km).

approximately linear relationship between the estimates and the straight-line (geodesic) distances between cities. The results shown in Figure 7(B) agree with this hypothesis for most of our estimates apart from some data points, including a few clearly erroneous ones. By manual inspection of the source data, we found out that the erroneous estimates are due to data irregularities: even after our data filtering pipeline some cell towers seem to suddenly change their location. Thus whenever the data itself is of reasonable quality, these two sanity checks demonstrate that our method yields sensible results.

To illustrate how well our estimates align with the real world, we consider two examples where travel time estimates are available from elsewhere. According to Lonely Planet [33], the travel time from St.-Louis to Dakar is roughly five hours with frequent *sept-place* taxis. Our peak estimate of 5 h 7 min matches this extremely well. Furthermore, if we look at our estimate of the travel time from Dakar to Ziguinchor (also discussed in Figure 2(C)) equals 14 h 30 min, which matches the approximate 15 h travel time to travel from Dakar to Ziguinchor by ferry [34].

3.2 Comparison with Google's estimates

To compare our estimates with existing routing engines, we also obtained travel time estimates between all city coordinates from Google's Distance Matrix API [35] using the default parameters of the service. The comparison of our and Google's estimates is shown in Figure 8. Overall, our results and those obtained through Google's API are roughly linearly dependent. Compared to our estimates, Google's estimates tend to be lower, especially when longer travel times are considered suggesting that Google may effectively overestimate the typical travel speed in Senegal. However, this is difficult to verify, as Google has not made public how they produce their travel time estimates. It is nevertheless clear that the baseline for Google's estimates originates from the road network data that includes information on road network types and speed limits. In many developed countries, Google is known to also track and store location data from the users of its mobile operating system [36]. This however requires that the people in the monitored area have

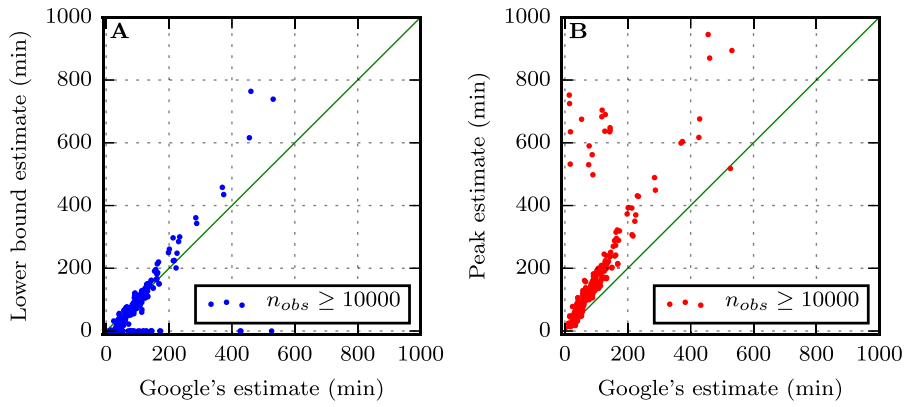

Figure 8 Comparison of our estimates with Google's Distance Matrix API estimates. Only estimates based on at least 10,000 inter-observation times are shown. In Panel **A**, we show our results for the lower bound estimate, and in Panel **B**, we show our results for the peak estimate. Google's estimates tend to be overall smaller than ours. This is most evident with longer distances, and indicates that Google seems to systematically overestimate travel speeds in Senegal.

access to smartphones, so that *e.g.* GPS location data can be transferred from the phone to Google. In Senegal, smartphone penetration is still low (15% of adult users, 2014 [18]), which can make it challenging for Google to calibrate their travel time estimates. Naturally, the differences between our and Google's travel time estimates can also originate from biases in our source data due to artifacts such as erroneous cell tower coordinates, varying coverage ranges of cell towers, and the non-continuous tracking of individuals. Also the goal for Google's estimates may be different than ours: Google may focus on providing estimates the travel time between places without including any additional delays *e.g.* for breaks. Thus, due to the lack of established ground-truth data, we can not claim either of the method to be superior - all that is certain is that there is a systematic difference.

3.3 Travel speed maps help to pinpoint anomalous travel times

To demonstrate how our results could be used for monitoring travel conditions in a country, we compute the speed of travel between Dakar and other Senegalese cities assuming that the travel follows straight lines and that the travel times equal the peak estimates. This we visualize in Figure 9(A). The results show in general that the further the distance from Dakar, the greater is also the speed of travel. As Dakar is known for its congestion this result is in line with expectations, although possible systematic biases due to *e.g.* smoothing of the distribution can affect the results.

Nevertheless, we can also find an exception to the rule: The travel speed from Dakar to Ziguinchor is arguably slower than to many other cities that are of same distance from Dakar. This is most likely due to the ferry travel between these two cities, as land travel requires one to cross Gambia and travel on bad roads.

3.4 Monitoring travel times: case study on opening of the Dakar-Diamniadio toll highway

Our method also allows near-real-time monitoring of travel times: all that needs to be done is periodically feeding CDRs to the peak detection algorithm - say, daily or weekly. This allows maintaining up-to-date travel time estimates, and in particular, detecting expected or unexpected changes in travel times.

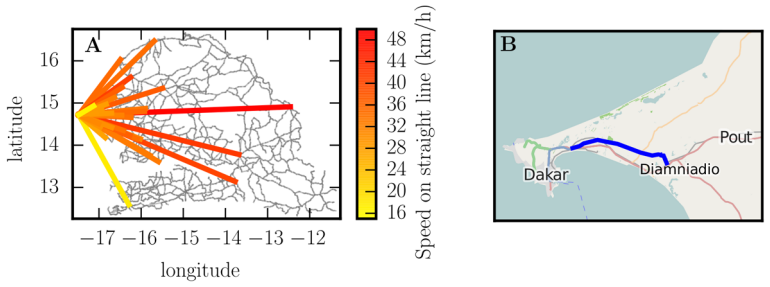

Figure 9 Getting out of Dakar takes time. On the left (Panel **A**), we display straight-line travel speeds from Dakar to selected other cities in Senegal. The figure shows that the longer the distance from Dakar, the faster the speed of travel, except for Ziguinchor to which travel often takes place by ferry resulting in slow travel speed. The grey network in the background represents Senegal's road network obtained from Ref. [37]. The gap in the road network between the southern and northern parts of Senegal corresponds to Gambia for which road network data is not included. On the right (Panel **B**) we show a close-up map showing the main roads close to the capital Dakar that is located on a peninsula. The stretch of the Dakar-Diamniadio highway that was opened on August 1st 2013 is highlighted with a thick dark blue line. The map image in Panel B is a modified excerpt from OpenStreetMap (© OpenStreetMap contributors).

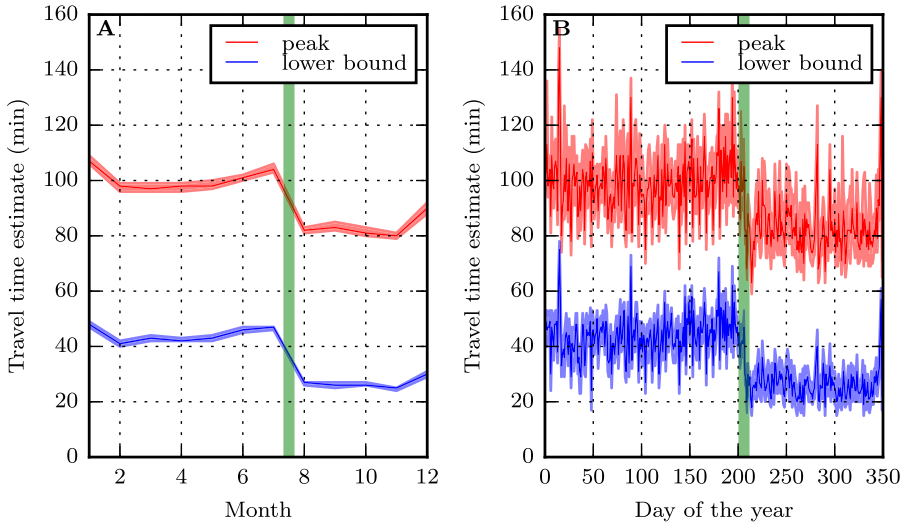

Figure 10 Opening of the Dakar-Diamniadio toll highway shortens travel times between Dakar and Pout. Daily (Panel **A**) and monthly (Panel **B**) travel time estimates between Dakar and Pout. The solid lines denote the estimated values, and the shaded areas denote the 5th and 95th percentiles of 1,000 bootstrap estimates. The green vertical line marks the opening date of the highway in both plots.

To demonstrate the potential of the method for this task, we have analyzed the decrease of the typical travel time between Dakar and Pout when the last part of the Dakar-Diamniadio toll highway was opened on August 1st 2013 [38]. The locations of the new highway stretch, Dakar, and Pout are shown in Figure 9(B).

In Figure 10 we show monthly and daily travel time estimates from Dakar to Pout (both peaks and lower bounds). From the monthly and daily estimates it becomes clear that there is a drop of 15 to 20 minutes in the typical travel time around the time when the new highway was opened. Not surprisingly, the results for the daily estimates are noisier than the monthly ones as they are based on fewer inter-observation times. Nevertheless, the

drop in travel times can be pinpointed with high accuracy to match the opening day of the highway.

4 Discussion

In this paper we have introduced a method for extracting typical travel times between cities from CDR-based mobility data. To demonstrate the usefulness of the method, we have applied it to data from Senegal, released for the 2014 D4D Challenge, and shown that it produces feasible estimates even though the spatial and temporal resolution of the data has been artificially reduced. Compared to Google's Distance Matrix API estimates, our approach yields estimates that are on average longer than those by Google suggesting the possibility that Google may be overestimating travel speed in Senegal. Also, we have discussed how the method can be used for monitoring changes in travel speeds in near real-time, as demonstrated by measuring the impact of opening a new highway on travel times.

For our method to work properly, a sufficient amount of data is required, especially when accurate travel time estimates are called for (for monitoring changes only, a higher level of noise is tolerable). However, even when only using a sample of 300,000 individuals out of Senegal's total population of over 14 million inhabitants, we were able to obtain reasonable travel time estimates between many Senegalese cities. Were the operator's CDR data to be used in full, we would have also been able to provide estimates for pairs of cities between which little traffic takes place.

Our method would also benefit from better spatial and temporal accuracy of data. An increased spatial resolution would allow better allocation of cell towers to cities, and if the temporal resolution of the data was improved, also within-city analyses based on individual cell-towers could become feasible. Note that for CDRs without artificial restrictions this is typically the case: positions of towers are accurately known, and data is recorded at a time resolution of one second. Further, were the data augmented with data on handovers, location area changes, and Internet usage data, the estimates would become even more accurate.

As some of our results point out, errors in data can give rise to corrupted results. While simple filtering of the data removed some of the errors, others did persist. Most likely these errors could have been avoided at source: the errors have to do with changing base station locations, base station ID's that have been switched between stations, or other technical issues at the operator's end.

In addition to improving the quality, amount, and accuracy of the source data, also the method itself could be tuned for more accurate estimates. For instance, one could make use of information on road network characteristics or distances between cities and use them to regularize the estimation problem *e.g.* using Bayesian methods. Moreover, the estimation process could also be approached in a more holistic manner using ideas originating from the triangle inequality: if we know the typical travel times t_{AB} and t_{BC}, it is likely that the typical travel time t_{AC} is smaller than or of the same order of magnitude as $t_{AB} + t_{BC}$. Thus, if there are few trips observed between cities A and C, the travel time $t_{AB} + t_{BC}$ could be used as a soft constraint for estimating the travel time t_{AC}. This approach could also help out in automatic detection of erroneous results that are due to irregularities in the source data.

It is also worth pointing out that the estimates produced by our approach are bound to suffer from different biases because of reasons ranging from varying cell tower ranges to

offset times between cell phone usage and traveling. The effect of these biases could be diminished with more accurate calibration data on human mobility, such as GPS location data recorded for a sample of users.

Finally, let us discuss certain benefits of the method. The method is easy to implement, and it does not require either massive deployments of sensors, GPS traces from smartphones, or large-scale computational resources as the analyses can be run even on a standard desktop computer. The method also avoids privacy concerns that are often associated with CDR data, as it can operate with chunks of anonymized data (all that is required is series of locations and times), and produces only aggregate data that does not violate the privacy of individual users.

The real value of any method comes nevertheless from its use in practice. For travelers, both locals or tourists, the information on typical travel times is valuable as it helps out in planning of trips. For transport infrastructure planners, the method provides means for spotting possible bottlenecks in a transportation network. Also, as the method lends itself to near real-time monitoring of travel conditions, it can be used for assessing changes in travel times, either locally or country-wide. Such changes can result locally from special events, deteriorated (or improved) road conditions, or disturbances such as illegal check point harassment. On a larger scale, one could envision detecting distruptions to travel patterns caused by disasters, violent conflicts, or outbreaks.

To summarize, in this work we have demonstrated that it is possible to extract and monitor travel times using CDR data when high quality data are available in sufficient amount. Given that the cost of applying the method in practice is low and the potential gains remain significant, we hope to see our method implemented also in practice - especially in developing countries where accurate travel time information is not often readily available.

Additional material

Additional file 1: Supporting web-site: Travel time estimates and inter-observation time distributions (zip)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors designed the research and wrote the paper. RK implemented the method and performed the numerical analyses.

Acknowledgements
Studies and researches were performed using mobile communication data made available by Sonatel and Orange within the D4D Challenge. Computing resources provided by the Aalto Science IT project are acknowledged. JS and TA acknowledge financial support from the Academy of Finland, project no. 260427. We thank Onerva Korhonen and Ilkka Kivimäki for comments on the manuscript.

Endnote
[a] Some of the positions of the base transceiver stations (BTS) that correspond to the cell tower IDs in the text are known to have changed locations during the time span of the CDR data. (Personal e-mail communication with the D4D-Senegal organizers, September 2014.)

Received: 21 October 2015 Accepted: 19 February 2016 Published online: 01 March 2016

References
1. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4:10
2. Naboulsi D, Fiore M, Ribot S, Stanica R (2015) Large-scale mobile traffic analysis: a survey. IEEE Commun Surv Tutor 18:124-161. doi:10.1109/COMST.2015.2491361
3. Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. Eur Phys J B 88:164

4. United Nations Global Pulse (2013) Mobile phone network data for development. http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev. Accessed 18 Dec 2015
5. Letouzé E (2014) What is big data, and could it transform development policy? PositionIT April/May:40-47. http://www.ee.co.za/article/big-data-transform-development-policy.html. Accessed 14 Dec 2014
6. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci USA 109:11576-11581
7. Blondel VD, Esch M, Chan C, Clérot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2012) Data for development: the D4D challenge on mobile phone data. Preprint. arXiv:1210.0137
8. de Montjoye Y-A, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD (2014) D4D-Senegal: the second mobile phone data for development challenge. Preprint. arXiv:1407.4885
9. Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in Rome. IEEE Trans Intell Transp Syst 12(1):141-151
10. Becker RA, Caceres R, Hanson K, Loh JM, Urbanek S, Varshavsky A, Volinsky C (2011) A tale of one city: using cellular network data for urban planning. IEEE Pervasive Comput 10(4):18-26
11. Berlingerio M, Calabrese F, Di Lorenzo G, Nair R, Sbodio ML (2014) AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast, pp 397-411
12. Angelakis V, Gundlegård D, Rydergren C, Rajna B, Vrotsou K, Carlsson R, Forgeat J, Hu TH, Liu EL, Moritz S et al (2013) Mobility modeling for transport efficiency: analysis of travel characteristics based on mobile phone data. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast, pp 412-422
13. Liu F, Janssens D, Cui J, Wets G (2015) Building workers' travel demand models based on mobile phone data. In: Data for development challenge Senegal: book of abstracts: scientific papers, pp 180-199
14. Wang Y (2015) Use of mobile phone data for planning a road network: application to the country of Senegal. Master's thesis, TU Delft
15. Coifman B (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transp Res, Part A, Policy Pract 36(4):351-364
16. Mori U, Mendiburu A, Álvarez M, Lozano JA (2015) A review of travel time estimation and forecasting for advanced traveller information systems. Transportmetrica A, Transp Sci 11(2):119-157
17. Cohen S, Christoforou Z (2015) Travel time estimation between loop detectors and FCD: a compatibility study on the Lille network. France Transp Res Proc 10:245-255
18. Poushter J, Oates R (2015) Cell phones in Africa: communication lifeline. http://www.pewglobal.org/files/2015/04/Pew-Research-Center-Africa-Cell-Phone-Report-FINAL-April-15-2015.pdf. Accessed 14 Dec 2014
19. Steenbruggen J, Borzacchiello MT, Nijkamp P, Scholten H (2013) Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal 78(2):223-243
20. Caceres N, Wideberg J, Benitez FG (2008) Review of traffic data estimations extracted from cellular networks. IET Intell Transp Syst 2(3):179-192
21. Rose G (2006) Mobile phones as traffic probes: practices, prospects and issues. Transp Rev 26(3):275-291
22. Qiu Z, Cheng P (2007) State of the art and practice: cellular probe technology applied in advanced traveler information system. In: 2007 TRB 86th annual meeting: conference recording. Transportation Research Board, Washington
23. Bar-Gera H (2007) Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel. Transp Res, Part C, Emerg Technol 15(6):380-391
24. Janecek A, Hummel KA, Valerio D, Ricciato F, Hlavacs H (2012) Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, New York, pp 361-370
25. Virtanen J (2002) Mobile phones as probes in travel time monitoring. Finnish Road Administration
26. Promnoi S, Tangamchit P, Pattara-Atikom W (2008) Road traffic estimation based on position and velocity of a cellular phone. In: 2008 8th international conference on ITS telecommunications. ITST 2008, pp 108-111
27. Wang H, Calabrese F, Di Lorenzo G, Ratti C (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 2010 13th international IEEE conference on intelligent transportation systems (ITSC), pp 318-323
28. Wunnawa SV, Yen K, Babij T, Zavaleta R, Romero R, Archilla C (2007) Travel time estimation using cell phones (TTECP) for highways and roadways. Final report for the Florida Department of Transportation
29. Innamaa S, Hätälä E (2012) Floating mobile data pilot in the Helsinki metropolitan area. VTT Technologies, Espoo. http://www.vtt.fi/inf/pdf/technology/2012/T51.pdf
30. de Montjoye Y-A, Quoidbach J, Robic F, Pentland AS (2013) Predicting personality using novel mobile phone-based metrics. In: Social computing, behavioral-cultural modeling and prediction. Springer, Berlin, pp 48-55
31. Tageo: Senegal city & town population. http://www.tageo.com/index-e-sg-cities-SN.htm. Accessed 19 Dec 2014
32. Torres C, Briceno-Garmendia C, Dominguez C (2011) Senegal's infrastructure: a continental perspective. World Bank Policy Research Working Paper Series
33. Ham A et al (2013) West Africa, 8th edn. Lonely Planet, London, p 339
34. Wikipedia, the free encyclopedia: MV Aline Sitoe Diatta. https://en.wikipedia.org/wiki/MV_Aline_Sitoe_Diatta. Accessed 30 Jul 2015
35. Google Distance Matrix API. https://developers.google.com/maps/documentation/distancematrix/intro. Accessed 20 Apr 2015
36. Quora: "Google Maps: how does Google Maps calculate your ETA?". https://www.quora.com/Google-Maps/How-does-Google-Maps-calculate-your-ETA. Accessed 14 Dec 2015
37. Senegal roads: Africa Infrastructure Knowledge Program. http://www.infrastructureafrica.org/library/doc/584/senegal-roads. Accessed 18 Dec 2014
38. Bank AD Senegal moves into the fast lane with the opening of its toll highway. http://www.afdb.org/en/news-and-events/article/senegal-moves-into-the-fast-lane-with-the-opening-of-its-toll-highway-12263. Accessed 16 Dec 2014