CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Collection Session 1
27-30 September 2021, Online 15 September 2021

Home and Work Identification Process Using Mobile Positioning Data

Nugroho, Amin Rois Sinung; Munaf, Alfatihah Reno M. N. S. P.; Madjida, Wa Ode Zuhayeni;Putra, Amanda Pratama Putra; Setyadi, Ignatius Aditya (Statistics Netherlands)

sinung@bps.go.id, alfa@bps.go.id, uhayeni@bps.go.id, amanda.putra@bps.go.id, adityasetyadi@bps.go.id

Abstract

The use of mobile phones has led to a huge growth of mobile positioning data (MPD). One of Indonesia's provider produces 7 to 10 billion records from around 190 million subscribers per day. The data comes from customer transaction data or Call Detail Record (CDR), and customer presence location or Location Based Service (LBA). The availability of this data provides an opportunity for the National Statistics Office (NSO) to use it as new data sources for official statistics such as tourism statistics and population mobility statistics.

Tourism is defined as activities that are conducted by a person outside of his/her usual environment. Population mobility is defined as activities that are conducted regularly by a person inside his/her usual environment. The usual environment usually consists of their home point and work point. Therefore, identifying their home and work points is a crucial first step. We have developed an approach/algorithm to identify a person's home and work points based on the person's movement data from MPD at city level accuracy. In this algorithm, we combined: the highest frequency of a person's location at night hours in a month, the highest frequency of a person's location at working hours during weekdays in a month, and duration spent in the home and work location to further increase the accuracy.

To test the accuracy of the algorithm, we conducted a preliminary validation using around 63 subscribers mobile positioning data in four months (December 2017 to March 2018). Our preliminary evaluation results with the ground truth showed that the accuracy of identifying home point is 96.23%.

Keywords:

Human Mobility; Mobile Positioning Data; MPD; Trip Identification; Staypoint

Home and Work Identification Process Using Mobile Positioning Data

Amin Rois Sinung Nugroho[1]; Alfatihah Reno M. N. S. P. Munaf[1]; Wa Ode Zuhayeni M[1]; Amanda Pratama Putra[1]; Ignatius Aditya Setyadi[1]

[1] BPS - Statistics Indonesia

Abstract:

The use of mobile phones has led to a huge growth of mobile positioning data (MPD). One of Indonesia's provider produces 7 to 10 billion records from around 190 million subscribers per day. The data comes from customer transaction data or Call Detail Record (CDR), and customer presence location or Location Based Service (LBA). The availability of this data provides an opportunity for the National Statistics Office (NSO) to use it as new data sources for official statistics such as tourism statistics and population mobility statistics.

Tourism is defined as activities that are conducted by a person outside of his/her usual environment. Population mobility is defined as activities that are conducted regularly by a person inside his/her usual environment. The usual environment usually consists of their home point and work point. Therefore, identifying their home and work points is a crucial first step. We have developed an approach/algorithm to identify a person's home and work points based on the person's movement data from MPD at city level accuracy. In this algorithm, we combined: the highest frequency of a person's location at night hours in a month, the highest frequency of a person's location at working hours during weekdays in a month, and duration spent in the home and work location to further increase the accuracy.

To test the accuracy of the algorithm, we conducted a preliminary validation using around 63 subscribers mobile positioning data in four months (December 2017 to March 2018). Our preliminary evaluation results with the ground truth showed that the accuracy of identifying home point is 96.23%.

Keywords:

Human Mobility; Mobile Positioning Data; MPD; Trip Identification; Staypoint

1. Introduction:

During this transformation phase in the relationships between global change and local development nowadays, human mobility is one of the main themes of study in geography. Government policies related to spatial planning and transportation usually require this data for analysis. Another benefit of understanding human mobility is reducing infectious diseases (Wesolowski, A., et al, 2012; Dalziel, B. D., et al, 2013), and disaster management (Lu, X., et al, 2012). Previously, to infer this movement is by doing the direct measurement, for example by counting the number of passengers departing and arriving at public transportation facilities (train stations, airports, terminals, etc.), the number of vehicles passing on the highway, or through surveys.

Today, most people cannot be separated from their mobile phones. The use of mobile phones is increasing year by year. Based on national socio-economic survey data conducted by BPS - Statistics Indonesia, in 2017 59.59% of the population in Indonesia owned a mobile phone, around a 2.21 percent increase compared to the previous year. By using a mobile phone wherever the person is located, they constantly create mobile positioning data with spatiotemporal dimensions namely geolocation sequence along with time stamps. The detected geolocation sequences can generate a travel diary of each mobile phone user.

Travel refers to the activity of someone who moves between different geographic locations for any purpose and any duration (United Nations, 2010). Group of travel to various places produces a trip. Based on the United Nations (2010), a trip is defined as the travel by a person from the time of departure from his usual environment until he returns. Trip determination is the core of human mobility algorithm, before being used for the analysis of tourism, transportation, migration, and so on (United Nations, 2010; Amin, I., et al, 2017; Dewulf, Bart., et al, 2017; Batran, Mohamed., et al, 2018). Two basic things used in trip determination is the identification of home and work location.

Our goal is to develop an algorithm to identify home and work location of the mobile phone users based on their mobile phone footprint. The home and work location predicted by the algorithm is then validated to 63 mobile phone users to get the accuracy of the algorithm.

2. Data and Methodology:

There are two types of mobile positioning data: active and passive. Active mobile positioning data: mobile positioning system or mobile tracing data in which the location of the mobile phone is determined with a permit and a special query using a radio wave (in app-based positioning). Passive mobile positioning data: automatically stored in the log files in mobile network operators (MNOs), such as billing memory; hand-over between network cells, Home Location Register, etc (Ahas, R., et al., 2008).

Passive mobile positioning data is used to achieve the objectives in this paper. The methodology might differ depending on the type of passive data provided by MNOs. The basic type of passive data by MNOs is the dataset of Call Detail Records (CDR) that represents phone activity — calls, messaging. Alternative data types include Data Detail Records (e.g. internet usage, SMS, call log), location updates (periodical identification of a device within the network antennae), and others. These data sources are more detailed and accurate in terms of frequency of the events and geographic accuracy, but they are often not available or require enhancement of the technology by MNOs (Eurostat, 2014). Location Based Service (LBS) is another passive mobile positioning dataset which will be generated due to the presence of cellular customers at a certain point. In this study, we use both CDR and LBS dataset.

msisdn charact	datetime timestamp without tir	source_dat character va	bts_lat character varyi	bts_lon character vary	province_name character varying (50	kabupaten_name character varying	kecamatan_name character varying (5	kelurahan_name character varying (5	node character	trx_date date
6281...	2018-02-08 13:33:03	LBA_ALL	-4.54691	120.35833	SULAWESI SELATAN	BONE	TANETE RIATTA...	CELLU	3G	2018-02-08
6281...	2018-02-08 13:51:39	CHG_POST	-4.54051	120.30777	SULAWESI SELATAN	BONE	TANETE RIATTA...	MACANANG	3G	2018-02-08
6281...	2018-02-08 14:00:33	CHG_POST	-4.53931	120.30337	SULAWESI SELATAN	BONE	TANETE RIATTA...	MACANANG	3G	2018-02-08
6281...	2018-02-08 14:03:08	CHG_POST	-4.54051	120.30777	SULAWESI SELATAN	BONE	TANETE RIATTA...	MACANANG	3G	2018-02-08
6281...	2018-02-08 14:17:06	LBA_ALL	-4.55271	120.38343	SULAWESI SELATAN	BONE	TANETE RIATTA...	BAJOE	3G	2018-02-08
6281...	2018-02-08 14:27:40	CHG_POST	-4.53961	120.33201	SULAWESI SELATAN	BONE	TANETE RIATTA...	MANURUNGNGE	3G	2018-02-08
6281...	2018-02-08 14:46:12	LBA_ALL	-4.54691	120.35833	SULAWESI SELATAN	BONE	TANETE RIATTA...	CELLU	3G	2018-02-08
6281...	2018-02-08 14:52:22	LBA_ALL	-4.54101	120.34101	SULAWESI SELATAN	BONE	TANETE RIATTA...	TA	3G	2018-02-08
6281...	2018-02-08 14:52:23	LBA_ALL	-4.54101	120.34101	SULAWESI SELATAN	BONE	TANETE RIATTA...	TA	3G	2018-02-08
6281...	2018-02-08 14:52:59	LBA_ALL	-4.55991	120.33282	SULAWESI SELATAN	BONE	TANETE RIATTA...	BIRU	3G	2018-02-08
6281...	2018-02-08 14:53:45	LBA_ALL	-4.55271	120.33401	SULAWESI SELATAN	BONE	TANETE RIATTA...	MASUMPU	3G	2018-02-08
6281...	2018-02-08 14:55:09	LBA_ALL	-4.55271	120.33401	SULAWESI SELATAN	BONE	TANETE RIATTA...	MASUMPU	3G	2018-02-08

Figure 1. Mobile phone footprint sample

We collaborated with Telkomsel which is the biggest mobile network operator in Indonesia with around 60% market share. Telkomsel mobile phone footprint data contains hashed MSISDN number of the subscriber, the transaction timestamp, the type of data stored, the BTS coordinates where the data is collected, the reverse geocode of the BTS coordinates, and also the network node which transaction is obtained. The data also includes the home and work identification by MNOs.

Since the data contains administrative areas, we performed matching processes with BPS's Local Administrative Unit (LAU) data before performing the analysis. BPS - Statistics Indonesia is updating the administrative areas regularly in 2 (two) periods each year (BPS, 2019). Telkomsel only provides local administrative units information by name, therefore we performed name-matching from the highest level (province) to the lowest (village) with BPS data.

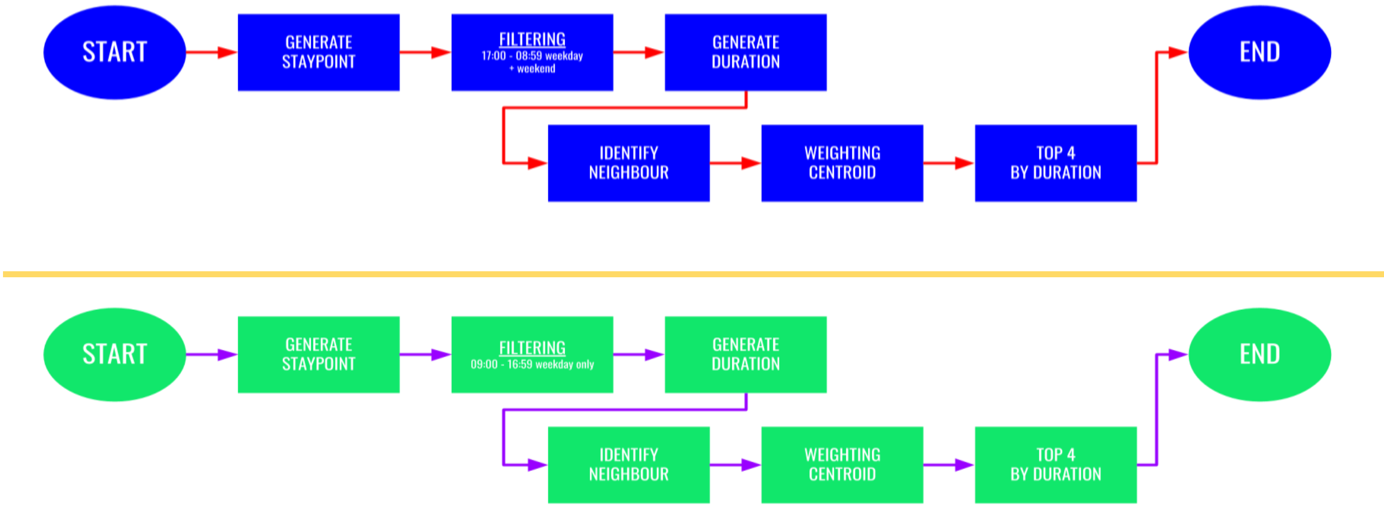

Figure 2. Home and Work Identification Algorithm

After the location administrative unit matching, we proceed with running the identification algorithm of home and work location of the subscribers as can be seen in Figure 2. Following are the steps of the algorithm in detail:

Step 1. Generate Staypoint.
Passive mobile positioning data (CDR and LBS) are sorted by timestamp for each subscriber. The ordered dataset is then referred to as trajectory. Staypoint is a grouping of the trajectory output. It clusters location updates which are close in space and time and calculates the centroid of that cluster to form a new location position as the staypoint of that cluster. (Dang, The Anh., Et al, 2017) provide a staypoint computation algorithm.

Step 2. Filtering.
- To infer home location, from the given staypoints, we take staypoint where the datetime is between 17.00 - 8.59 (nighttime) on weekday and all staypoint detected on the weekend.
- to infer work location, from the given staypoints, we use stay point in weekdays and the time is between 09.00 - 16.59 (working hours).

Step 3. Generate Duration.
In this step, we sum up the duration for every stay in the same exact location (longitude, latitude).

Step 4. Identify Neighbour.
This process calculates the distance in meters and finds the closest stay point with distance < 1 km.

Step 5. Weighting Centroid.
Generate new latitude and longitude as an average (weighted by duration). Iterate until there is no possibility to generate a new staypoint.

Step 6. Sorting.
New latitude and longitude are then sorted by duration, and selected as the top home and work location candidate.

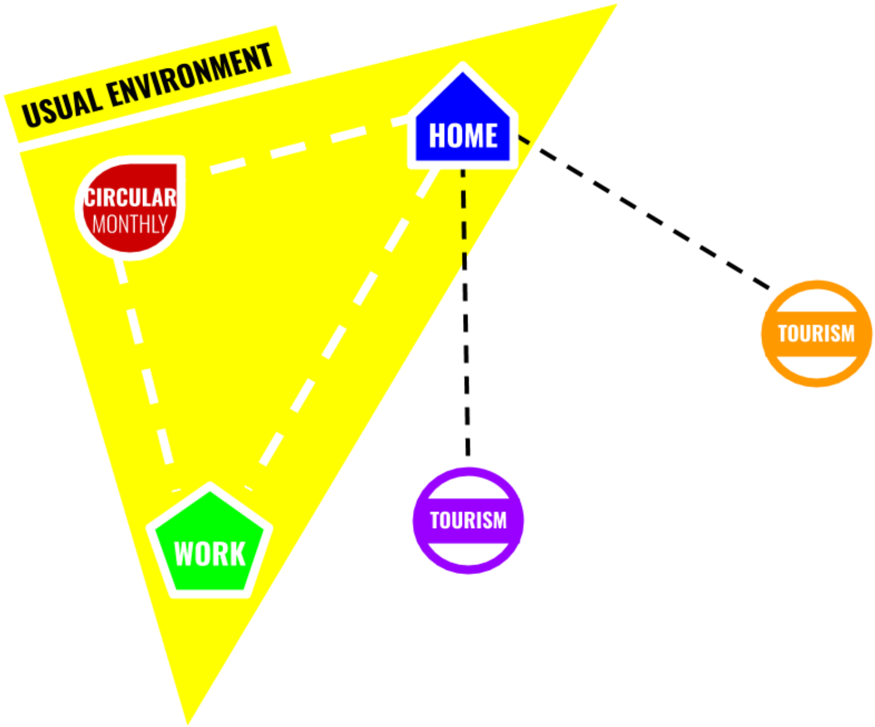

Figure 3. Usual environment model

BPS and Telkomsel agreed to develop trip identification algorithm and other statistics together. The next step is to predict the usual environment, so from every subscriber, we will have their usual environment and trips outside the usual environment (Figure 3). To check the accuracy of the algorithm prediction, BPS recruited volunteers who agree to give their Telkomsel based location data to be monitored and used for the ground truth that will be cross-checked with the algorithm prediction. As mobile positioning data is prone to affect a person's privacy, these 60 subscribers are BPS employees who has traveled outside of their usual environment at city level during these four months. They volunteered to support this study and have signed a consent form, following Telkomsel data privacy regulation, to allow their mobile positioning data to be used in this study. After applying the algorithm, we asked the volunteers to validate the result as the ground truth to be compared with our algorithm. We then applied the algorithm to the whole Telkomsel subscribers dataset after several iterations on these volunteers dataset until we are quite confidence with the accuracy.

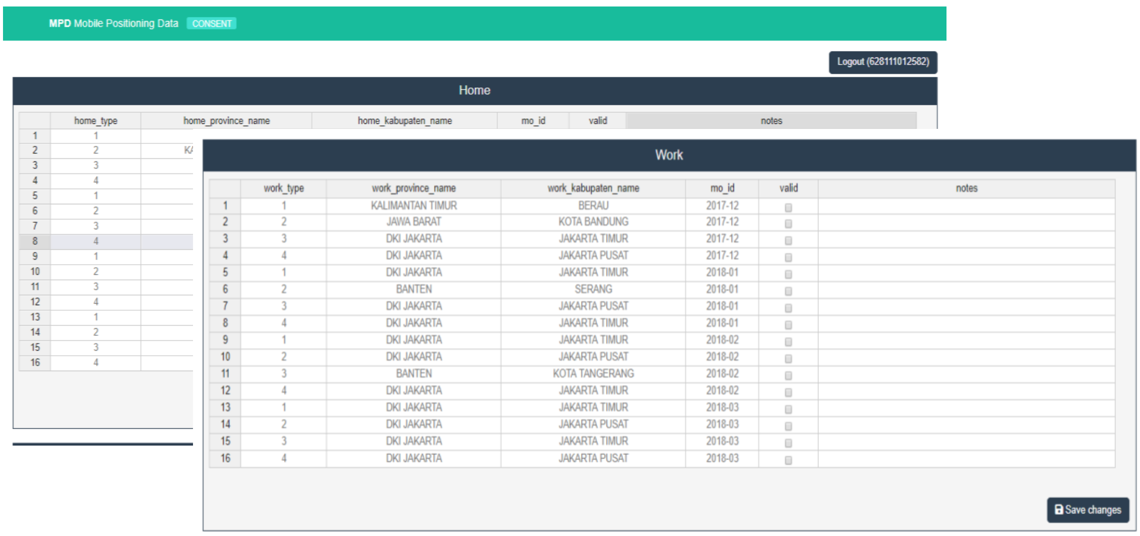

Figure 4. Volunteer validation main page

3. Result:

Indonesia has several Local Administration Unit (LAU) stages, starting with RT and RW at the lowest level and Kelurahan (Village), Kecamatan (District), Kabupaten/Kota (Regency/City), and Province at the top level. After running the algorithm, we asked the volunteers to verify/validate the home and work location prediction results at all of those LAU stages. The results showed that at lower LAU level, the prediction accuracy was around 58% which is not good enough. However, the accuracy is quite good at Kabupaten/Kota level which is above 80%. Consequently, we decided to apply the algorithm only at Kabupaten/Kota level.

We further proceed with testing several scenarios to increase the accuracy at Kabupaten/Kota level. After exploring the dataset, we predict the usual location using duration and days spent as weighting factor and only taking the top 1 candidates of home and work location from each month will increase the accuracy. To validate this prediction, we calculated and compared the accuracy of the following scenarios:

The first scenario is taking the top 4 (four) candidates of both home and work locations in every month, then extract the only top 4 (four) from the whole six months candidate locations. We applied this method on Kabupaten/Kota levels with and without using the duration and days spent as weighting factor. The accuracy result without the weighting factor turns out to be 75.47%, whereas the result with duration and days spent used as weighting factor turns out to be to 94.33% which is very good.

In the second scenario we only consider 1 (one) home and work candidate, the top one on every month from the whole candidate locations for six months. The same method with and without using duration and days as weighting factor is then applied. It showed the same pattern in which the one using duration and days as weighting factor also performed better. Furthermore, combining it with only taking the top 1 (one) home and work location candidate further increased the accuracy. The accuracy for the one that does not apply the duration and days spent as weighting factor increases to 90.57% and the one that uses duration and days as weighting factor produces more promising accuracy which is 96.23%. Based on this result, we are confidence to apply the algorithm with this method to the whole Telkomsel subscribers data.

The results of applying home and work identification algorithm to the whole dataset are then validated by comparing the number of subscribers based on their predicted home region with BPS official population projections data for each region. The number of subscribers was previously calibrated to be compared later with traditional survey. Figure 5 shows the distribution of calibration rate between Home location from MPD and the one from population projection data.

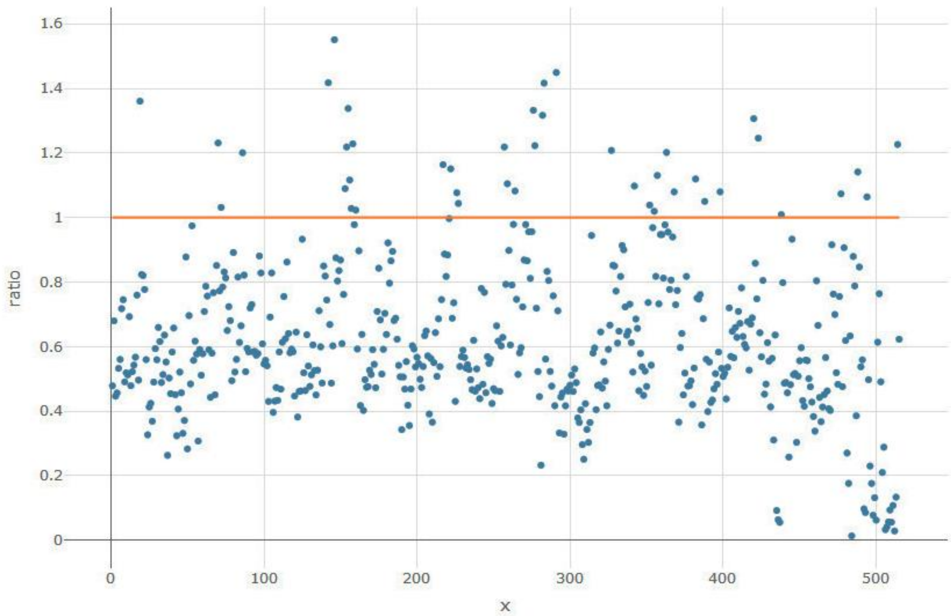

Figure 5. Distribution of Calibration Ratio between Home from MPD and Population Projection

4. Discussion and Conclusion:

This paper showed how to use passive mobile positioning data to infer human mobility. The determination of the home and work algorithm has been proposed and can be used later in trip determination. The accuracy of the proposed algorithm reaches 96.23%, but this accuracy is only for algorithms implementation at Kabupaten/Kota (Regency/City) level. Calibration with population projection can be an initial step to be used in comparing the results of mobile positioning data with conventional surveys. Our next plan for this research is to extend the algorithm development and the utilization in producing official statistics, for example tourism statistics, commuting statistics, and circular mobility statistics.

Since then, we have developed another algorithm for determining home and work location in 2019. Anchor Mobility Data Analytic (AMDA) is proposed for Home-Work Location Determination from Mobile Positioning Data, where this algorithm uses clockwise reversal to make it easier to classify someone in their usual environment. Only about 80% of the raw data can be used for the establishment of ordinary environments, the remaining 20% do not have sufficient data history. In this study, it was found that the accuracy of AMDA in determining monthly home location was 98.8% at the provincial level, and 88.7% at the regency level. As for the determination of monthly work locations, 98.9% at the provincial level, and 70.4% at the regency level.

References:

Ahas, R., et al., (2008). Evaluating passive mobile positioning data for tourism surveys: an Estonian case study. https://doi.org/10.1016/j.tourman.2007.05.014

Amin, I., et al. (2017). Inferring Commuting Statistics in Greater Jakarta from Social Media Locational Information from Mobile Devices.

Batran, Mohamed., et al. (2018). Sensing Population Mobility through City Boundary in Greater Maputo via Mobile Phone Big Data Mining. DOI: 10.1109/SERVICES.2018.00018.

BPS - Statistics Indonesia. (2019). Sistem Pemutakhiran MFD dan MBS Periode Pemutakhiran Semester 1 Tahun 2019. http://mfdonline.bps.go.id/.

Dalziel, B. D., et al. (2013). Human Mobility Patterns Predict Divergent Epidemic Dynamics Among Cities. DOI: 10.1098/rspb.2013.0763

Dang, The Anh., et l. (2017). Mobility Genome - A Framework for Mobility Intelligence from Large-Scale Spatio-Temporal Data. DOI: 10.1109/DSAA.2017.20

Dewulf, Bart., et al (2015). Examining Commuting Patterns using Floating Car Data and Circular Statistics: Exploring the use of new methods and visualizations to study travel times. http://dx.doi.org/10.1016/j.jtrangeo.2015.08.006.

Eurostat. (2014). Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics.

Lu, X., et al. (2012). Predictability of Population Displacement after the 2010 Haiti Earthquake. DOI: 10.1073/pnas.1203882109.

United Nations. (2010). International Recommendations for Tourism Statistics 2008. New York. United Nations Publication.

Wesolowski, A. et al. (2012). Quantifying the Impact of Human Mobility on Malaria. DOI: 10.1126/science.1223467