Zagatti, Guilherme Augusto et al.

Article

A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR

Development Engineering

Provided in Cooperation with:
Elsevier

Suggested Citation: Zagatti, Guilherme Augusto et al. (2018) : A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR, Development Engineering, ISSN 2352-7285, Elsevier, Amsterdam, Vol. 3, pp. 133-165, https://doi.org/10.1016/j.deveng.2018.03.002

Development Engineering 3 (2018) 133–165

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR

Guilherme Augusto Zagatti [a,*], Miguel Gonzalez [a], Paolo Avner [b], Nancy Lozano-Gracia [b], Christopher J. Brooks [a], Maximilian Albert [a], Jonathan Gray [a], Sarah Elizabeth Antos [b], Priya Burci [a], Elisabeth zu Erbach-Schoenberg [a,e], Andrew J. Tatem [a,e], Erik Wetter [a,d,f], Linus Bengtsson [a,c]

[a] Flowminder Foundation, Roslagsgatan 17, Stockholm, SE 11355, Sweden
[b] The World Bank Group, 1818 H St NW, Washington, DC 20433, USA
[c] Karolinska Institute, Department of Public Health Sciences, Stockholm, SE 17177, Sweden
[d] Stockholm School of Economics, Stockholm, SE 11383, Sweden
[e] WorldPop, Geography and Environment, University of Southampton, University Road, Southampton, SO17 1BJ, UK
[f] Asian Institute of Management, 1260 Makati City, Manila, Philippines

ARTICLE INFO

Keywords:
Urbanisation
Call detail records
CDR
Non-supervised learning
Commuting
Urban planning

ABSTRACT

The rapid, unplanned urbanisation in Haiti creates a series of urban mobility challenges which can contribute to job market fragmentation and decrease the quality of life in the city. Data on population and job distributions, and on home-work commuting patterns in major urban centres are scarce. The most recent census took place in 2003 and events such as the 2010 earthquake have caused major redistributions of the population. In this data scarce context, our work takes advantage of nationwide de-identified Call Detail Records (CDR) from the main mobile operator in the country to investigate night and daytime populations densities and commuting patterns. We use a non-supervised learning algorithm to identify meaningful locations for individuals. These locations are then labelled according to a scoring criteria. The labelled locations are distributed in a grid with cells measuring 500 × 500 m in order to aggregate the individual level data and to create origin-destination matrices of weighted connections between home and work locations. The results suggest that labor markets are fragmented in Haiti. The two main urban centres, Port-au-Prince and Cap-Haïtien suffer from low employment accessibility as measured by the percentage of the population that travels beyond their identified home cluster (1 km radius) during the day. The data from the origin-destination matrices suggest that only 42 and 40 percent of the population are considered to be commuters in Port-au-Prince and Cap-Haïtien respectively.

1. Introduction

Rapid urbanisation in Haiti with the absence of economic growth has led to increasing socioeconomic challenges. Urban areas have shown steady population growth fueled by both migration and natural growth. In 2015, official statistics suggest 53 percent of the total population was considered urban in Haiti (IHSI, 2015). With almost 6 million Haitians living in urban areas, cities now host over 0.5 million more inhabitants than rural areas. The rapid, unplanned urbanisation in Haiti creates a series of urban mobility challenges which can contribute to job market fragmentation and decrease the quality of life in the city. Data on population distributions and home-work commuting patterns in major urban centres is scarce (Prud'homme and Kopp, 2011). This paper aims to inform the debate about challenges brought by rapid urbanisation in Haiti by focusing on identifying commuting patterns in Port-au-Prince and Cap-Haïtien, the main metropolitan areas of the country hosting about 3.5 and 0.5 million inhabitants each. The evaluation of connectivity and employment patterns can uncover the

* Corresponding author.
E-mail addresses: guilherme.zagatti@flowminder.org (G.A. Zagatti), miguel.gonzalez@flowminder.org (M. Gonzalez), pavner@worldbank.org (P. Avner), nlozano@worldbank.org (N. Lozano-Gracia), chris.brooks@flowminder.org (C.J. Brooks), maximilian.albert@flowminder.org (M. Albert), jonathan.gray@flowminder.org (J. Gray), santos1@worldbankgroup.org (S.E. Antos), priya.burci@flowminder.org (P. Burci), elisabeth.zu.erbach@flowminder.org (E. zu Erbach-Schoenberg), andy.tatem@flowminder.org (A.J. Tatem), erik.wetter@flowminder.org (E. Wetter), linus.bengtsson@flowminder.org (L. Bengtsson).

https://doi.org/10.1016/j.deveng.2018.03.002
Received 18 October 2017; Received in revised form 9 March 2018; Accepted 19 March 2018
Available online 21 March 2018

extent of spatial mismatch in cities, and point at priorities for policy intervention.

Haiti faces challenges in undertaking traditional estimations of commuting patterns from survey data. Since the last national census was undertaken in 2003, Haiti has suffered a series of tragedies including an earthquake in 2010, a cholera outbreak in 2011 and Hurricane Matthew in 2016. Such events had a significant impact on the population distribution and urbanisation trends, and hence lead to questions regarding the reliability of existing population projections based on past trends. Origin-Destination (OD) surveys have been successfully used to identify commuting statistics in Nairobi, Kenya JICA (2013) and Buenos Aires, Argentina (PTUMA, 2010). However, such surveys can be expensive and time-consuming, thus reducing periodicity and prevalence. Furthermore, Haiti has never conducted either a comprehensive OD survey nor an economic census.

While there is no real substitute for the detail and precision of a full OD survey designed to be representative of all moves within a city, the country is well-suited for a study based on mobile phone data. Haiti's leading mobile network operator has over two-thirds of the market share in mobile phone subscriptions in the country (CONATEL, 2016). Information on location, time, and volume of phone calls made through mobile devices can provide valuable insights into where people live and work, and their patterns of movement over time. In this study we gathered and analysed data from mobile phone Call Detail Records (CDRs), with the aim of leveraging the mobility information contained in CDRs in order to examine commuting patterns. The methods presented in this paper build on previous research that has successfully used CDRs to estimate meaningful location for individuals.

The rest of this paper is organised as follows: Section 2 provides an overview of related work, Section 3 presents the data used in this study, Section 4 explains the proposed modelling approach. Main outputs are presented in Section 5. The paper concludes with a brief discussion in Section 6.

2. Related work

Several studies have made use of CDR data to study human mobility, both on individual level using disaggregated data (González et al., 2008; Song et al., 2010; Schneider et al., 2013; Lu et al., 2013; Pappalardo et al., 2015) as well as on population level, aggregating CDRs (Kujala et al., 2016; Tatem et al., 2009; Calabrese et al., 2011). More specifically, some studies have attempted to derive general laws about individual mobility by estimating future locations based on historical trends and the level of predictability of human mobility. These studies have shown that human mobility can be highly predictable (González et al., 2008; Song et al., 2010; Lu et al., 2013; Pappalardo et al., 2015). In other studies, CDRs have been used to study aggregated commuting patterns. Validation of the predicted flows through travel surveys or traffic counts has highlighted the potential of these data for estimating commuting patterns in contexts where other data is not available (Graells-Garrido and Saez-Trumper, 2016; Iqbal et al., 2014; Järv et al., 2012).

Other studies have used CDRs successfully to identify meaningful locations in people's lives on a finer resolution (Ratti et al., 2006; Pulselli et al., 2008; Reades et al., 2009; Ahas et al., 2010) and have also studied flows between those places (Isaacman et al., 2011). Ahas et al. (2010) surveyed the geographic research literature on meaningful places, which are also known as "anchor points", "term bases" or "core stops" in related literature. Meaningful places are defined as "places where individuals usually spend a considerable amount of time and which they consider important in the conduct of their everyday lives; these are typically the home and workplace" (Ahas et al., 2010). The researchers devised an algorithm which identified meaningful places of users in Estonia at the tower level. Using a different and simpler, algorithm, (Isaacman et al., 2011) identified meaningful places of users in Los Angeles and New York. The identification of meaningful places focuses on individuals and their relationship with space. Since the focus of this paper is on connectivity issues and the spatial mismatch between home and work locations, we chose to investigate the meaningful locations of individuals.

In order to identify meaningful locations of individuals in both Port-au-Prince and Cap-Haïtien and ascertain likely commuting behaviours from identified home and workplace locations, this paper takes an approach similar to that of Isaacman et al. (2011). This approach allows us to achieve a finer spatial resolution than the base station level by exploiting the high temporal resolution of the data. For scaling we investigate the methods presented in Deville et al. (2014).

3. Data

The datasets available for analysis included CDRs covering the period from 1st March to 30th May 2016 as well as antennae location data (Base Station IDs, see Appendix A). There were approximately 2 billion recorded events in the CDR dataset, with an event corresponding to a call or SMS transactions. Each event in the CDR contained the originating and destination IMSI (de-identified), MSISDN (de-identified), IMEI (de-identified), TAC, base station ID and international prefix as well as the starting time, duration and type of transaction. During a typical weekday in Haiti, the number of active users increases exponentially beginning at around 4:00, reaches a plateau between 9:00 and 17:00, and then peaks 40 percent above the plateau between 18:00 and 20:00. After peaking, the number of active users decreases rapidly, being close to zero at around 3:00. (A detailed description of the normal operation of a mobile phone network including the meaning of standard abbreviations and how CDRs are captured by the network can be found in Appendix A).

The Base Station ID dataset contained the base station ID, the latitude and longitude of the base station and the base station generation (e.g., 1G, 2G, 3G, 4G). We approximated base station coverage using a Voronoi tessellation which assigned for each tower (the set of co-located base stations) location a polygon which contains all points in the plane for which such tower is the closest one. The Voronoi tessellation approach assumes that a cell phone would connect to the closest tower and that tower coverages do not overlap. Individual tower ranges were estimated as the intersection between the Voronoi tessellation and the national network coverage provided by the mobile network operator.

Tower density — tower per square kilometre — can vary quite significantly across Haiti. In more populated and urban areas tower density is high. The Port-au-Prince Metropolitan Area has the highest tower concentration in the country, reaching densities of more than six towers per square kilometre followed by Cap-Haïtien which has less than a third of the tower density of the national capital. In the Port-au-Prince Metropolitan Area itself, there is significant variation in tower density across its large and varied urban space. The highest tower densities are found in the main business districts, where they can be more than double the tower densities seen in adjacent areas and five to ten times those densities in the outskirt.

The variability in base station range has implications for the use of CDR data for population density estimation since network transactions are only captured at the base station level. In some regions of Haiti, such as the Northwest, average base station coverage can reach up to 90 square kilometres. The average coverage tends to decrease around more populated and urban areas. In Port-au-Prince the average base station coverage is below 10 square kilometres. The average base station coverage decreases the closer one gets to the centre of the metropolitan region. Turgeau and Saint Martin have an average coverage below the equivalent of a circle with a diameter of 900 m which is equivalent to roughly three street blocks. In more residential areas, the base station range increases significantly. In the outskirts of Port-au-Prince, the average base station range is up to 10 times larger than in the centre. In Cap-Haïtien the mean base station range in the centre is about twice

that in the centre of Port-au-Prince. Similarly to the national capital, base station range decreases the farther away one goes from the centre of Cap-Haïtien.

Once we combined the CDRs and the base station datasets, it was possible to obtain the location of the base station which a user was connected to when their SIM card became "active", that is, whenever an event was recorded in the database.

Finally, in order to scale the estimated evening cell phone user population to the total population, the 2016 population predictions produced by IHSI (2015) were used. These predictions are based on an outdated population census from 2003 which has been mostly updated using administrative records. Since 2003, a number of disasters have affected the country including an earthquake in 2010, a cholera outbreak in 2011 and Hurricane Matthew in 2016. Such events can have significant impacts on the population and urbanisation trends, limiting the reliability of population projections based on the 2003 census. Nevertheless, these were the most comprehensive population estimates available for Haiti at the time the analysis was undertaken.

4. Methods

4.1. Clustering meaningful locations

Compared to GPS, the accuracy with which we can determine the exact location of individuals is lower, as it depends on the density of mobile phone towers. However, in urban areas such as in Port-au-Prince and Cap-Haïtien tower density is high enough to warrant a good approximation of a user's location. A further limitation of CDRs is that the sampling mechanism is dependent on call frequency, meaning we only have a record of a user's location for the times he makes or receives a call. Since we define meaningful places as places where individuals spend most of their time, such as their home and workplace we are likely to accumulate enough samples when aggregating over a period of time. It is assumed that these meaningful places exert a gravitational effect on the users, such that the farther users are from any of their meaningful places, the less likely they are to spend time there. Further, we assume that calls are location heteroscedastic such that a user is more likely to place calls at their meaningful places because they tend to spend more time in there and might find it more convenient to place the majority of calls in those locations.

Therefore, the distribution of call events will be centred around those places which are meaningful for an individual. For instance, as in Fig. 1, if a user lived closer to the centre of a radio cell, a substantial proportion of call events would be captured by the radio cell which contains the user's home place. Adjacent base stations would capture fewer call events. As such, a good estimate for the user's home place would be the centroid of the radio cell. On the other hand, if the user lived close to the border of a radio cell then it would be more likely that the neighbouring radio cell would capture approximately the same volume of call events. In this case, the best estimate for a user's home would be a point mid-way between the two neighbouring radio cells.

However, if a user had their home and workplace in two neighbouring radio cells it would be possible to expect that both radio cells would capture a substantial level of call activity, but in this case it would not make sense to estimate a user's home or work location as a point between the two radio cells because it would just be conflating the two places in one estimate. Thus, it is useful to assume that the gravitational effect exerted by a given place has a limited radius of influence. Hence, it is only valid to estimate the location of a given place, say the home location, using radio cells which are within a limited distance apart from each other. An algorithm that captures the assumptions outlined above would cluster a number of towers together whose distance does not exceed a given threshold. The weighted centroid of each cluster can then be used as the best estimator of each one of a user's meaningful location.

The information contained in a user's CDR was used to produce accurate estimates of meaningful locations in Isaacman et al. (2011). The first step of their algorithm sorts the towers a user was connected to in terms of the days the tower was contacted ("call-days"). This is an important step in order to minimize the effect of a flurry of activities which are not associated with the presence of a user in one of their meaningful locations. For instance, trips which are short in duration might be associated with a spike in the number of calls, as a user might call back home to family and friends. These calls would unduly increase the perceived importance of the location. In the second step, the CDR events are clustered according to Hartigan's leader algorithm (Hartigan, 1975) as described in Isaacman et al. (2011). This algorithm has a deterministic outcome in the sense that it does not require initialisation with a pre-specified number of clusters or centroids. Moreover, the algorithm is quite efficient with large datasets, as in the present case.

Following Isaacman et al. (2011), Hartigan's leader algorithm is illustrated in Fig. 2 and proceeds as following:

1. Select a threshold.
2. Sort all the towers by the number of "call days" (i.e., the number of days a user was connected to a given tower).
3. Start from the tower with the highest number of call days and form a cluster with a centroid in this tower.
4. Move to the next tower:
   a) Descend through the cluster list in the order in which they were created.
   b) If the tower is located at a Euclidean distance less than the threshold from the centroid of the current cluster then include that point in the cluster and recompute the centroid of the cluster weighted by the number of calls.
   c) If the next tower does not comply with (a), then move to the next cluster and check condition (a). If there is no cluster for which the tower satisfies (a), then create a new cluster with a centroid in the current tower.

In the absence of validation data, it is very difficult to calibrate an unsupervised learning algorithm as the present one. This is a problem inherent to many other clustering algorithms such as k-means, hierarchical or Gaussian mixture clustering models. Within- and between-cluster variation metrics are usually used to determine the ideal parameters of each model. In the present case, we adopted a number of strategies to select the ideal threshold size.

The key variable in Hartigan's leader algorithm is the threshold size. As the size of the threshold increases the number of clusters tends to reduce as shown in Fig. 3. Since the Euclidean distance used in this exercise satisfies the triangle inequality, the number of clusters will always decrease or remain the same as the threshold size increases as proved by Hartigan (1975), but the rate at which this happen will depend on empirical factors. On the other hand, the mean and the standard deviation of distance between towers and cluster centroid will tend to increase as we increase the threshold size.

In Fig. 3 we ran the algorithm on a randomly selected sample of 10,000 users in Port-au-Prince using different threshold levels to assess the sensitivity of the results to the threshold chosen. As expected and shown in the left panel of Fig. 3, the number of clusters per user tends to decrease as the threshold size increases. The rate at which this happens substantially increases as the threshold size surpasses the closest tower mean distance and tapers off as the threshold reaches 10 km.

In Isaacman et al. (2011), the CDR data was validated using information provided by volunteers, who identified the precise location of key sites such as home and workplace. Based on these data, Isaacman et al. (2011) performed a logistic regression of the clusters identified on a number of derived CDR statistics such as the total number of clus-

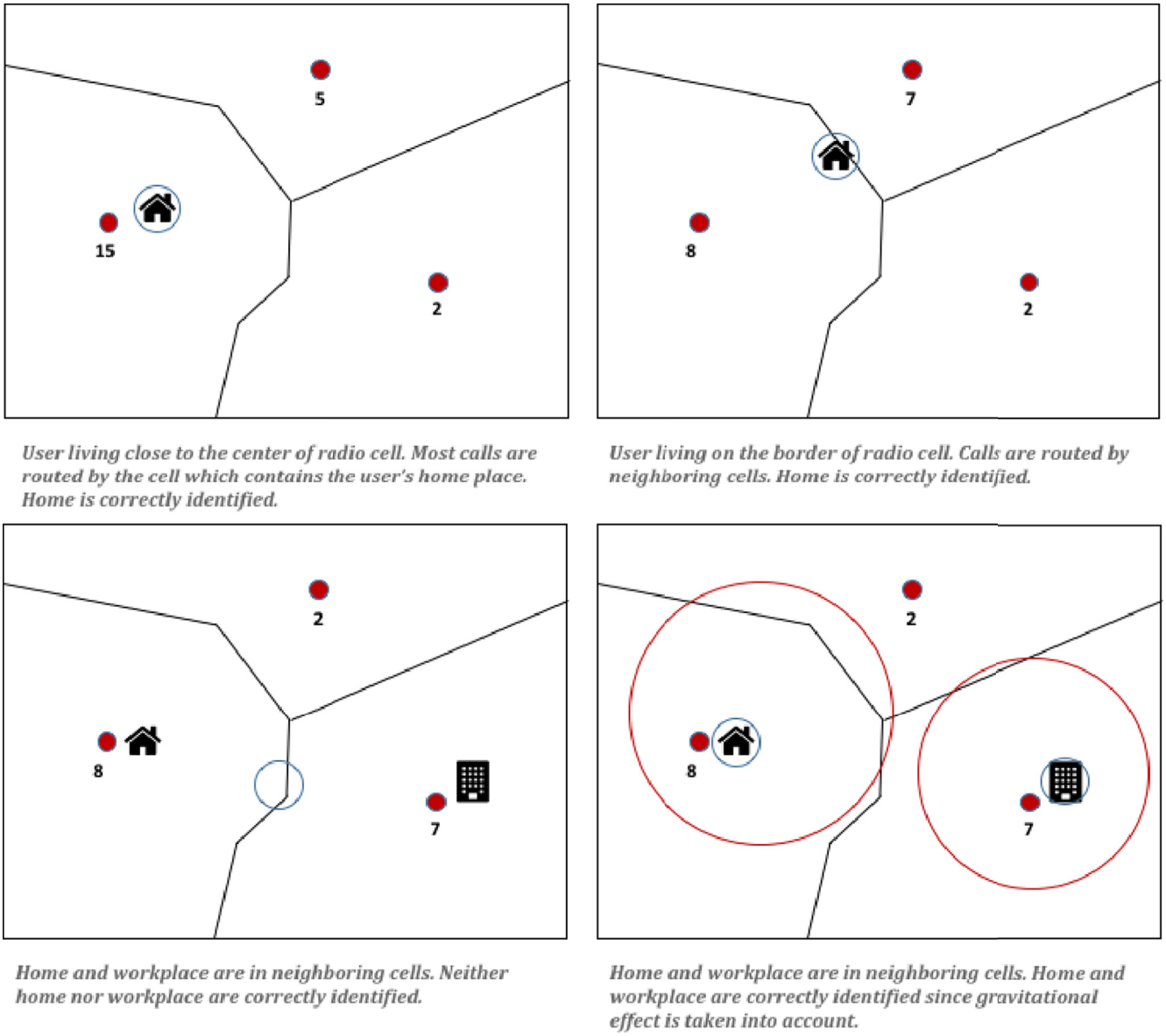


Fig. 1. Issues around identifying meaningful locations.

ter "call-days", the number of days between the first and last cluster "call-day", the number of times any tower in the cluster was contacted between 13:00 and 17:00 during weekdays and the number of times any tower in the cluster was contacted between 19:00 and 7:00 of the following day at any day of the week.

The researchers found that the most important features for identifying the "meaningfulness" of a cluster were the total number of "call-days" and the number of days between the first and last cluster "call-day". Those measures remove transient places from the list of identified clusters. As expected from the assumptions outlined above, the researchers also found that meaningful places from which a user generates few CDR events were not identified by the algorithm. In the validation phase, Isaacman et al. (2011) found that the approach maintained within-4.8 km-accuracy of the meaningful location for 88 percent of the users.

Since low "call-day" towers can potentially add undesirable noise to the data, all the clusters which do not contain an aggregated number of "call-days" above nine — which represents 10 percent of the total days in the study — have been filtered out in the present case.

Fig. 3 provides an overview of the results for the clustering exercise, when several alternative distance thresholds are applied. The right panel includes an additional constraint where all clusters with less than nine "call-days" are excluded from the analysis. As the threshold size increases, the number of users who have five or more clusters decreases substantially, leaving most of the users with a single cluster.

Isaacman et al. (2011), find that for cities like Los Angeles and New York, a distance of 1.6 km works well as a threshold for the cluster analysis.[1] Isaacman et al. (2011) report that in their target urban area towers might be as dense as 200 m apart, while in suburban areas spacings of 1.6–4.8 km are more common. For this reason, they choose a threshold of 1.6 km. In the centre of Port-au-Prince the median distance to the closest tower is 350 m while in suburban areas the median distance rises to about 500 m. The rate at which the mean and the standard deviation of the distance between towers and cluster centroid increase can provide some indication for the ideal threshold. Whereas the mean centroid distance will tend to shoot up as the cluster becomes more

[1] Work on the optimal selection of this threshold is beyond the scope of the present study.

Fig. 2. Hartigan-clustering algorithm.

heterogeneous, the standard deviation will increase at a slower rate as seen in Fig. 4 which depicts the coefficient of variation. At 1 km, the coefficient of variation stabilizes.

For the reasons above we chose a threshold of 1 km, which is a compromise between the likely catchment area of a meaningful location and the amount of towers potentially covered by it. However, further research is necessary to determine the optimal threshold.

4.2. Labelling meaningful locations

This section focuses describing the process for classifying the meaningful locations identified in the previous section according to the period of the day which individuals spend in those locations, with particular emphasis on the distinction between day and evening periods of the working week. The objective of this part of the exercise is to

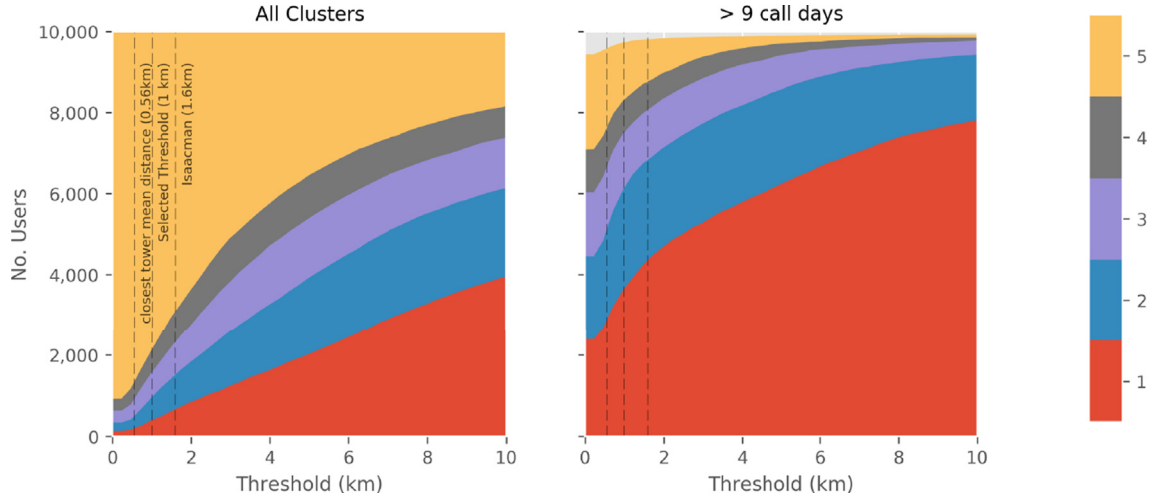


Fig. 3. Number of clusters per user.

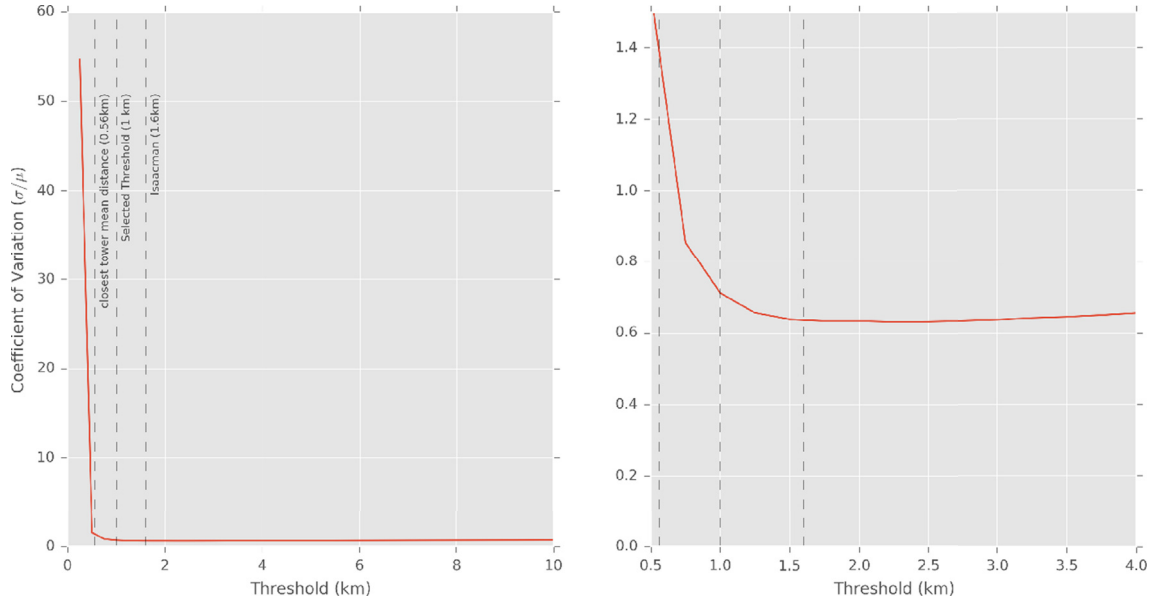


Fig. 4. Coefficient of variation of the distance between towers and cluster centroid.

identify home and work locations based on the location of individuals during day and evening periods.

Locations are meaningful for different reasons and those reasons are likely correlated with time. For instance, it is expected that users will sleep at home most days, thus spending most of their nights there, especially during working days when they might not be travelling or visiting friends and family. Another meaningful location might be the user's workplace where the user might spend most of their daytime during working days but not during weekends and holidays. As such it is reasonable to assume that the presence of the user at any of their meaningful locations is correlated with the time of the day and the day of the week.

In order to understand where individuals systematically spend their time, it is necessary to make assumptions about their routines. As a starting point, it is assumed that individuals are likely to place or receive calls with the same probability throughout their waking hours, which here are assumed to range from 7:00 to 23:00. Likewise, it is also assumed that individuals are likely to place calls with the same probability throughout the week. If we assume that call placement is location heteroscedastic, then call events can only be independent of time if the location of an individual is independent of time as well. Therefore, to label one of the meaningful places identified above as home or work using time information, we need to assume that call placement at different locations is not independent of time.

If location is dependent on time, then the call frequency distribution at a given location should be centred around the period in which a user is more likely to be present there. Let $p(c)$ be the probability that an individual makes a call at a given time and let $p(x)$ be the probability that the individual is at a given location x at a given time of the day. Then the probability that a call made in a given location at time t is, according to Bayes' rule: $p(t \mid c, x) = \frac{p(x|c,t)p(t|c)}{\sum_t p(x|c,t)}$.

Assuming that call placement is independent of time, we have that $p(t \mid c) = p(t)$ which is constant across time. Then the probability of observing many calls at a given period of the day is proportional to the amount of time spent in a given place, that is: $p(t \mid c, x) = \frac{p(x|t)p(x|c)p(t)}{p(x|c)\sum^t p(x|t)p(t)} = \frac{p(x|t)}{p(x)}$.

Therefore, the meaningful locations identified in the previous section can be classified according to the period of the day in which they are most active in terms of network activity. Since the distribution of meaningful locations can be characterised according to the period of

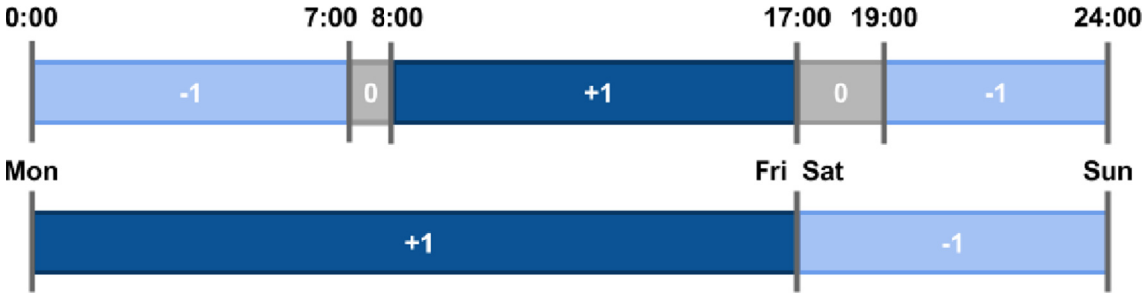


Fig. 5. Scoring criteria.

the day, such distributions can be used to estimate the population distribution during the day and evening times. Second, based on the identified commuting patterns between day and evening times, it is possible to develop an understanding of the location of residential and business districts.

Using ground-truth data from volunteers, Isaacman et al. (2011) classified the identified meaningful locations into home and work locations by means of two independent algorithms. The algorithms calculate a score for each important cluster using coefficients obtained from a logistic regression of the ground truth classification on a number of CDR derived statistics. The researchers found that the single most dominant factor for home classification was the rank of the number of events that take place in the evening. For work classification the first most dominant factor was the rank of the number of events which take place in the daytime followed by the inverse of the number of events which take place in the evening. This reflects the assumption that a person's workplace is somewhere where they do not usually spend their nights.

The researchers found that the home and work algorithms classified 50 percent of the clusters as home or work with errors below 1.5 km and 1.3 km respectively. Moving out to the 95th percentile the home and work algorithm achieved 6.1 km and 34.1 km errors respectively. The substantial increase in the reported error at the 95th percentile for the "work" algorithm was mainly caused by the volunteers not using their cell phone regularly at work.

In the present case, no ground truth data are available to run a logistic regression. As such, another approach which captures the assumptions outlined above and the logic present in Isaacman et al. (2011) is necessary. The clusters identified in the previous section are classified according to a scoring criteria. The scoring criteria assigns two-dimensional points to each event in the CDR dataset based on its timestamps. The first dimension — the hour score — captures the hours of the day. Events which fall between 7:00 and 8:00 and between 17:00 and 19:00 are assigned a score of zero. Events between 8:00 and 17:00 are assigned a score of one and events between 19:00 and 7:00 of the following day are assigned a score of minus one. The second dimension — the weekday score — captures the days of the week. Events which fall between Monday and Friday are assigned a score of one and events that fall in the weekends are assigned a score of minus one. The scoring criteria is illustrated in Fig. 5. Once all the events are assigned a score, the total score is aggregated by cluster and normalised by the total number of CDR events.

Analogously to Isaacman et al. (2011), the hour score weights calls during the day and evening time in opposite ways such that call events that fall during the night will contribute against call events which happen during the daytime. To soften this effect two buffers are constructed around the beginning of the day and at the end of the day, where call events do not add up to the score. Similarly, the day score will assign positive scores to call events that fall during the weekday and negative

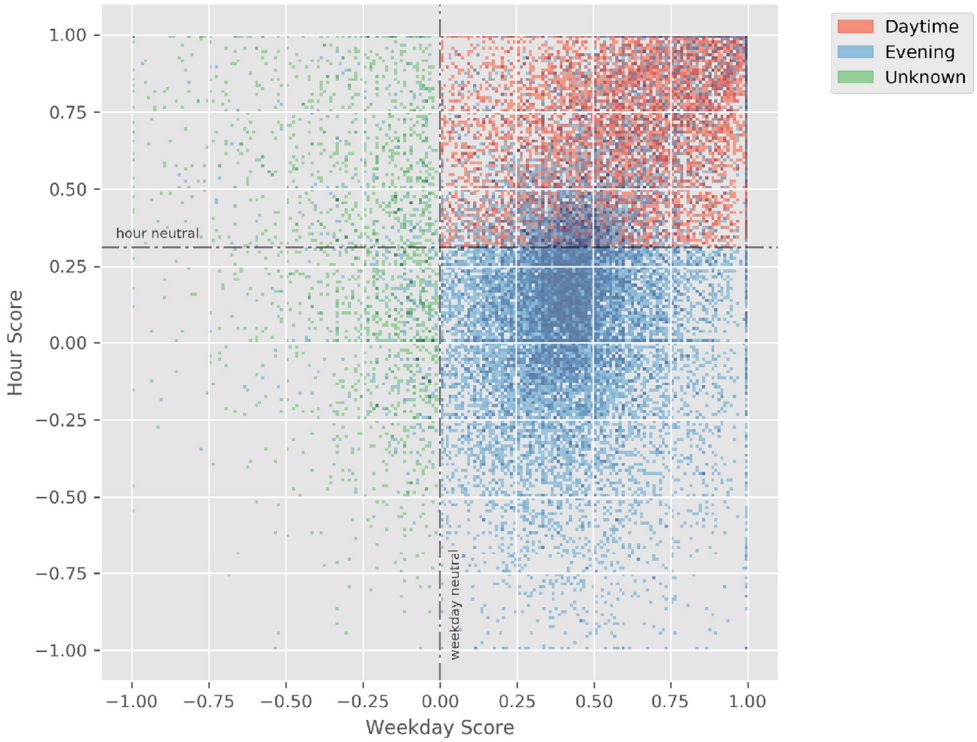


Fig. 6. Hour score versus weekday score.

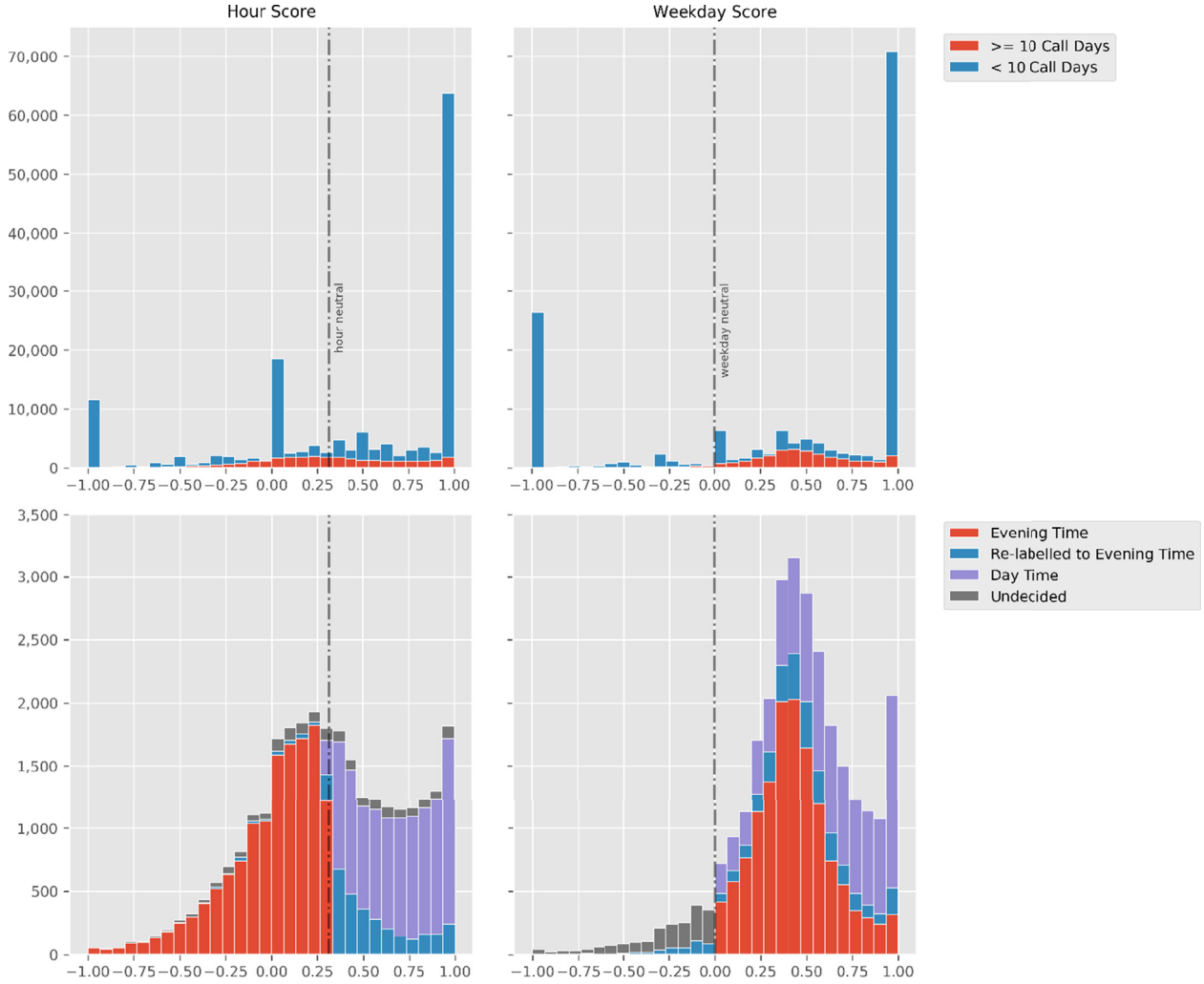


Fig. 7. Distribution of hour and weekday scores.

scores to calls that fall in the weekend.

In order to classify the clusters into day and evening time clusters, the neutral score is calculated. The neutral score represents the score that would be assigned to a cluster if CDR events were evenly distributed across time in that cluster. Assuming that a user is usually awake between 7:00 and 23:00, the neutral hour score would be 0.3125. On the other hand, assuming that a user places calls with equal probability throughout the week, the neutral weekday score would be 0.429. Given that many call events take place on weekends and that some users might work Saturdays and/or Sundays, the neutral weekday score is then adjusted downwards and set equal to 0. Further research is required to determine the ideal neutral scores for classification purposes.

Clusters which have an hour score above the neutral time score and a day score above the neutral weekday score are classified as daytime clusters. On the other hand, clusters which have an hour score below the neutral hour score and a weekday score above the neutral day score are classified as evening time clusters. All the other clusters are left undecided. Once all clusters are classified, a second classification procedure is followed.

Since the scoring criteria do not ensure that all users will be assigned an evening cluster, all the clusters of those users who do not have an evening cluster are re-labelled as evening. This choice should bias the share of commuters in the total population downwards, but avoids the opposite and stronger effect that would arise from dropping those individuals without a meaningful evening location. This step is taken under the assumption that all the users in the dataset must have an evening location. Further, for those individuals whose clusters conflate home and work locations, their call patterns do not likely have a strong signature since their calls are likely evenly spread during their waking hours. Those are the individuals whom the algorithm might not be able to identify a night-time meaningful place for. Figs. 6 and 7 show that most of them are indeed located close to the neutral lines. If we were to drop those observations in which the algorithm cannot determine a night-time meaningful place, the share of non-commuters in the total population would decrease, but the absolute number of commuters and its spatial distribution would remain the same (ignoring those few individuals who have an unknown and a daytime clusters). That means that commuting distance statistics across commuters would remain unaffected.

The methods do not guarantee that all individuals will have day and evening time locations. Those individuals who have both locations are considered commuters. Those who do not have locations in both categories are seen as non-commuters and both their day and evening time locations are assumed to be the same. Those are individuals who might work close to home or who might not work at all and for whom the call signature is random.

Second, the algorithm above does not ensure that all users will have a single day or evening time cluster. For instance, it is possible that some users have two evening time clusters. In those cases, each cluster will be counted proportionally to the total number of CDR events such that all clusters with the same classification contribute a total of one to

the total population count.

In Fig. 6 the scoring algorithm was run on the randomly selected sample of 10,000 users in Port-au-Prince as in the previous section. The figure depicts all the clusters found by running Hartigan's leader algorithm across the scoring space. The upper right-hand corner above the hour neutral score and right to the weekday neutral score represents all the clusters classified as daytime clusters. On the other hand, the lower right-hand corner below the hour neutral score and right to the weekday neutral score represents all the clusters classified as evening time clusters. All the clusters on the left hand side are left as undecided. The clusters which have been re-labelled as evening clusters can be spotted as those clusters outside of the lower-right quadrant shaded in blue. Most of them are in the bottom of the daytime cloud.

The distribution of clusters is depicted in the histograms in Fig. 7 which shows cluster counts across the hour and weekday score ranges. The first row depicts all the clusters found by running Hartigan's leader algorithm before filtering. A significant amount of noise can be seen when clusters which have less than ten call days are included as they tend to cluster on the extremes of the scoring range.

The second row depicts only the clusters which have not been filtered out. For this particular sample, the algorithm found that all the users had an evening location, of which 20 percent were re-labelled as a result of the scoring algorithm, 47 percent had a daytime location and 13 percent had some clusters left undecided.

The panel on the left shows the histogram for the hour score. Before re-labelling, a little more than half of the clusters are classified as evening time clusters and the rest as daytime. After re-labelling, about 30 percent of clusters to the right of the hour neutral line are re-assigned to evening clusters. The distribution of clusters to the left and right of the hour neutral line is quite different. Whereas the number of clusters tend to decrease as one move further to the left of the hour neutral line, the number of clusters is quite uniform to the right of the hour neutral line, even peaking at the far right. That is indicative of the fact that evening clusters have their call pattern more evenly distributed across the day. Whereas, daytime clusters have their call pattern largely concentrated during daytime, suggesting that those places might actually be workplaces.

When looking at the panel on the right, it is possible to note that most of the evening clusters are located around value 0.429, which represents the true weekday neutral score as discussed above implying that calls are randomly distributed across the week. On the other hand, daytime clusters scores is skewed to the right, suggesting that call pattern is more prevalent during weekdays constituting another indication that those might indeed be workplaces.

In the absence of ground truth data, it is impossible to validate this algorithm. Nevertheless, we performed sensitivity analysis to assess how robust the final results are to changes in key parameters. In order to carry out the sensitivity analysis, those parameters are varied within certain bounds and the resulting outputs are compared with a view to identifying systematic variations which could invalidate the final results. The results from the sensitivity analysis suggest that there is no systematic variation from the final choice of labelling parameters. More information can be found in Appendix B. Finally, it is worth mentioning that the method does impose a minimum commute distance since a person who works and live within a 1 km radius will be labelled as non-commuter. The very definition of commuting is subjective. The objective of this paper was to consider only individuals who commute above a certain threshold, below which commuting would be much harder to detect and would not be as relevant to policy-making.

4.3. Griding meaningful locations

Hartigan leader's algorithm can only produce point estimates associated with high uncertainty given that the performance of the algorithm decreases as the base station range increases. In order to minimize this effect, the point estimate is turned into an area estimate and an equal probability is assigned to each point in the estimate.

The area estimate is constructed as a buffer of 750 m around the point estimate if the cluster contains more than one tower, as is often the case in urban areas. When the point estimate only contains one tower, the cluster centroid is equal to the coordinates of the cell phone tower. In such case, the area estimate is chosen as the Voronoi polygon encompassing that tower. Home-work commuting link from a point in a home buffer to a point in a work buffer will count for the probability mass in the home buffer times the probability mass in the work buffer. If a person does not commute, we only spread the probability mass around the estimated point. We do not estimate commutes between points in the buffer.

The buffer size was chosen after experimenting with different sizes, and 750 m proved to be large enough to remove border effects but small enough to minimize the loss of resolution. This number can be contrasted with the standard errors reported by Isaacman et al. (2011) described above. Further discussion can be found in Appendix C.

The estimates are then intersected with a Universal Transverse Mercator zone 18N regular grid containing grid cells measuring 500 m by 500 m. The mass emanating from a grid cell equals the area of overlap between the area estimate and the cell. For instance, if half of the estimate falls in a given cell, that cell will count for half of a location. Each grid cell is summed across all individuals in order to obtain the population distribution. The counts are proportionally adjusted according to the area of intersection between the area estimate and the grid cell.

4.4. Scaling

The methodology described in the previous sections allows one to estimate the evening cell phone user population. However, the goal of this paper is to understand the distribution and dynamics of the entire Haitian population. Therefore a scaling procedure which provides a mapping from cell phone to population numbers is required.

The final estimated evening cell phone user population density can be seen in Fig. 8. This final estimate was obtained by running the algorithm described in the previous sections on the whole CDR dataset. The dataset contained approximately 5.2 million users, 83 percent of which were considered for analysis since they had at least one cluster above nine call-days. A total of 10.8 million clusters — both day and evening time — were found, which implies 2.5 clusters per user on average. The highest concentration of cell phone users is located in and around Port-au-Prince. The sprawl around the national capital is considerably larger than elsewhere in the country. Following Port-au-Prince, one finds large urban sprawls around Cap-Haïtien, Gonaïves and Les Cayes, all of which are regional capitals.

In order to scale the estimated evening cell phone user population to the total population, we considered two models. The first model was a simple linear model $\rho_c = \gamma + \alpha\sigma_c$, where ρ_c is the 2015 population density at administrative 3 (or "Commune") level predicted from the census by IHSI (2015) which was chosen as the target variable since the population census likely reflects evening population given focus on residence. γ is a constant and σ_c the evening cell phone density. To merge the target and covariate datasets, we sum all the covariate grid cells that fall within the target geographic region. Those cells that fall at the border of geographic regions are weighted by the percentage of the overlapping area. This method of aggregation is similar to the one used in Tatem (2017). We run the regression at the highest target resolution we have to scale our estimates. The linear model assumes that beyond a population density baseline, cell phone density is proportional to population density. For instance, one could assume that mobile phone ownership is a constant factor across households.

The second model was a super-linear model $\ln\rho_c = \ln\alpha + \beta\ln\sigma_c$, which assumes that $\gamma = 0$, that is, population density baseline is zero. On the other hand, the model assumes that the elasticity of cell phone user density to population density is not unit, that is, a percentage

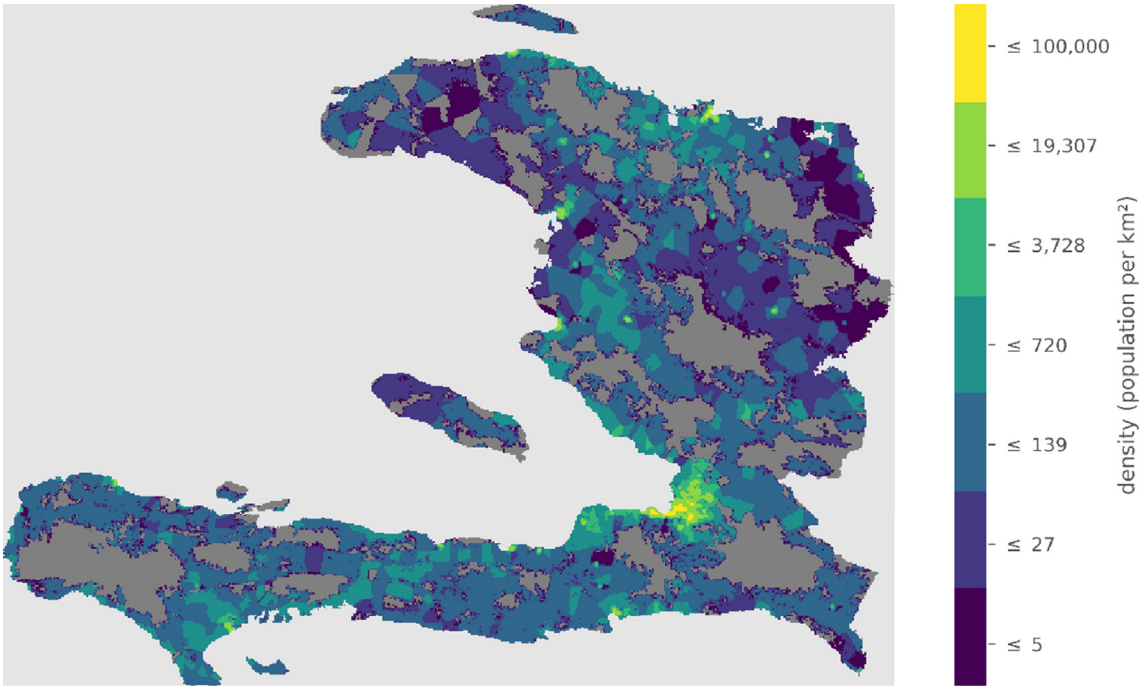


Fig. 8. Evening time cluster distribution.

change in cell phone ownership does not translate in a constant percentage change in population density. It could be argued that regions with higher cell phone density would be regions where marginal increases of cell phone users would not translate in significant marginal population density increases since in those regions cell phone ownership would be closer to population saturation, as is often the case in urban areas in high income countries. Alternatively, cell phone ownership might be highly concentrated such that an increase in cell phone ownership does not necessarily reflect significant population density increases. Table 1 shows the coefficients of the models estimated with OLS procedures.

The linear model produces a much stronger fit to the data as can be gleaned from the lower standard error. The standard error from the linear model is 1.5 times the mean population density which is about half of the standard error from the super-linear model. The linear model indicates that population density is about 1.7 times larger than the cell phone user density beyond a baseline of 180 people per square-kilometre.

On the other hand, the super-linear model suggests that the elasticity of cell phone user density to population density is rather small at 0.391 ± 0.02. Deville et al. (2014) use a super-linear model to produce dynamic population maps for Portugal and France, where they find elasticities of 0.803 ± 0.015 and 0.902 0.036 for each country respectively. The researchers report that the results are rather sensitive to the values of β with significant increases in the root-mean square errors within different scenarios analysed.

Despite the methodological difference in estimating cell phone user density between Deville et al. (2014) and this paper, the significant difference of the elasticity between Haiti (0.391), Portugal (0.803) and France (0.902) could be explained by the fact that cell phone penetration is much lower and uneven in Haiti, which is a developing country with a significant proportion of the population living in poverty. A 1 percent increase in cell phone user density in Haiti probably translates into a small percentage increase in population density, since higher user density in a particular area, such as a richer area, could be due to higher penetration rates, but not necessarily higher population densities.

At the same time, it is possible to argue that the scale ratio (α) is much larger in Haiti than in Portugal and France. Deville et al. (2014) points out that changes in α are corrected by total population adjustments. In Haiti a phone user is likely to account for many more people than a phone user in Portugal or France, because there are few cell phone users relative to the total population.

Despite the arguments in favour of the super-linear model, the model performance is considerably poorer than the linear model. Fig. 9 depicts the fit for both models. The linear model has a confidence interval which encompasses a much larger number of points, especially closer to the extremes of the distribution. When compar-

Table 1
Linear and super-linear model for scaling cluster distribution. Significance at 0.05 (*) and 0.001 (**).

	Linear model			**Super-linear model**		
No. obs.	570			562		
R-squared	0.823			0.413		
Log-like.	−4772.90			−185.99		
Std. error (absolute)	1048.80			2299.30		
Std. error ($\overline{\rho_c}$)	153.421			335.341		
	Coef.	**Std. error**	**z**	**Coef.**	**Std. error**	**z**
γ	179.1501**	45.054	3.976	0	–	–
α	1.6748**	0.033	51.333	58.2580**	1.08893	48.128
β	1	–	–	0.3911**	0.02	19.831

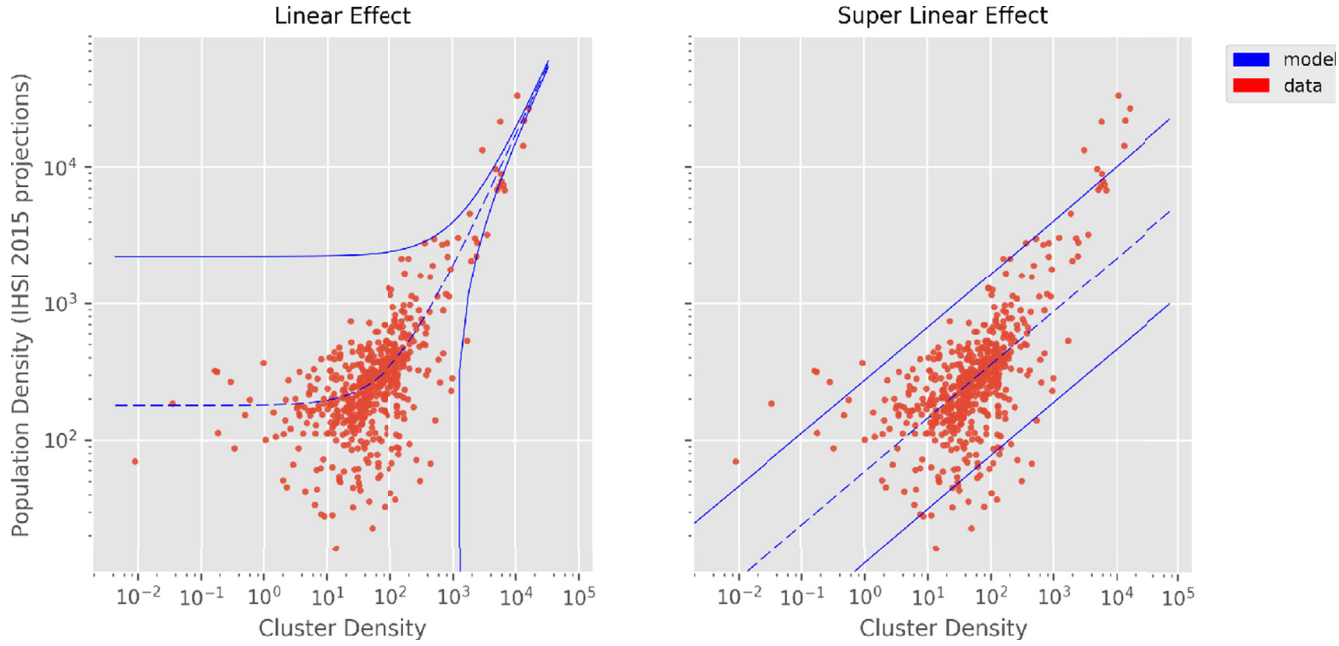


Fig. 9. Linear versus super-linear model fit – log scale.

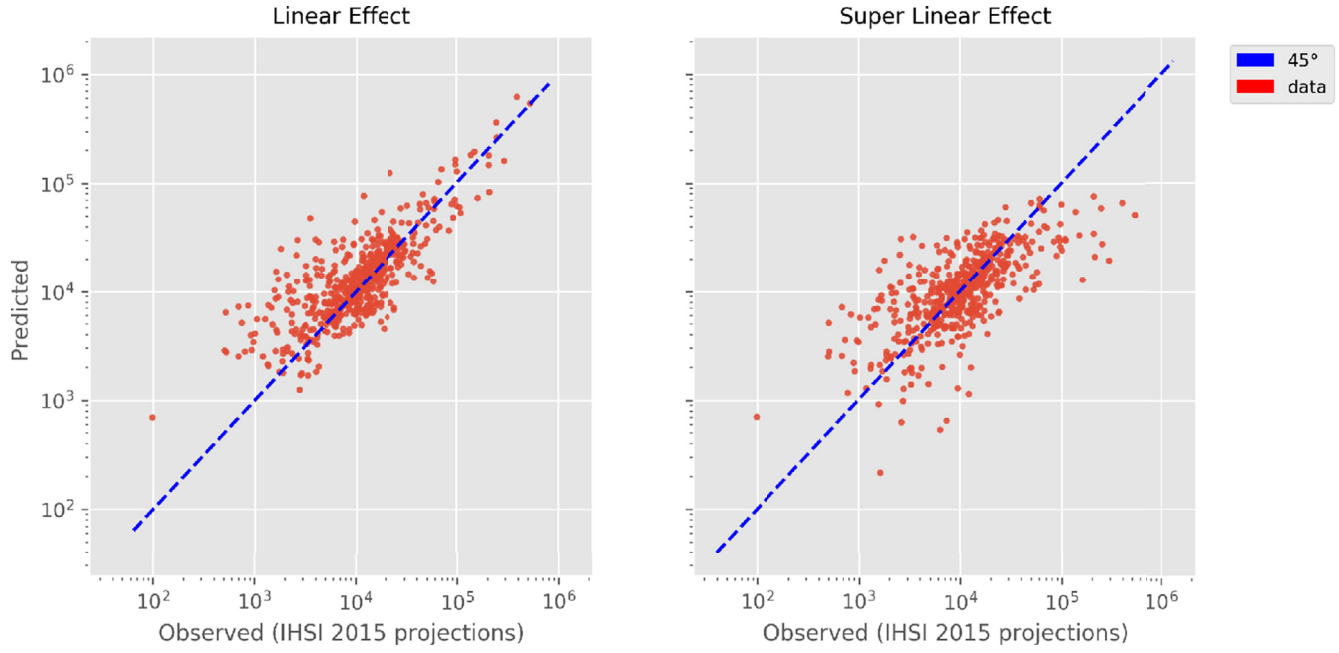


Fig. 10. Observed versus predicted population density — log scale.

ing the predicted values from the model with the IHSI (2015) projection values in Fig. 10, it is possible to confirm that the linear model produces an improved fit around the whole range of the distribution.

Given the improved performance of the linear model, this model was chosen over the super-linear model to scale the evening cell phone user density. Further, the linear model renders itself to simpler and more intuitive interpretations while providing an improved fit to the data. This model is much more flexible in accounting for disparate penetration rates than the log-log model which holds the intercept, γ, equals to null. By allowing the intercept to vary, the linear model is able to account for a mean baseline population density which does not have access to cell phones at all.

The predicted population density using the linear model is depicted in Fig. 11. Regions without any evening clusters were set to zero under the assumption that most regions without any cluster would be de-populated, especially around metropolitan areas which are the focus of this paper. Nevertheless, it is a fair critique to say that some vast areas of the country which are predicted to be empty might have people which are not captured by the methodology described in the previous section. Those areas are mostly located in the countryside of Haiti where mobile phone ownership and usage could be expected to be rather low. In those areas, the model is expected to perform much worse and make relative comparisons between rural areas much less accurate. In urban areas which is the focus of the paper, the model performs much better and relative comparisons are more

accurate.

Since the focus of this report is on the metropolitan regions of Port-au-Prince and Cap-Haïtien, it is important to assess the model fit for these two areas. In the first row of Fig. 12, predicted and IHSI (2015) projection values are compared. In Port-au-Prince it can be seen that a number of areas were overestimated especially towards the middle of the distribution. The second row of Fig. 12 depicts the ratio between predicted and observed population figures. Port-au-Prince is characterised by regions on its outskirts where the model produces relatively high over-predictions and areas in the centre where the opposite is true and under-prediction is more common. Although that may be caused by flaws in the linear model, there is another competing reason why that might be the case.

In the case of Port-au-Prince, the 2010 earthquake destroyed a significant number of structures in the city centre causing many businesses to move from the centre to the outskirts, especially towards Pétion-Ville in the Southeast. The earthquake also promoted the development of Canaan in the North fringe of Port-au-Prince, which served as a temporary camp to those affected by the disaster and developed into a sprawling slum (Nöel, 2012). Other disasters which affected rural areas of the country like hurricane Mathew in 2016 that affected the southern coast have likely caused a number of Haitians to migrate to Port-au-Prince and settle in the fringes of the town where land is more available. Given this anecdotal evidence, it is possible to assume that the changes spurred by these events would have caused the sorts of changes predicted by the model and not captured by IHSI (2015) population projections based on the outdated 2003 census. Thus, one would see under-prediction in the centre and over-prediction in the periphery if comparing the model with the outdated census.

On the other hand, these sort of events would have caused little impact to Cap-Haïtien, in which case the IHSI (2015) population projections would be closer to reality and the predicted model would be closer to the IHSI (2015) projections. This is indeed what can be seen from the scatter plot comparing observed and predicted values in Cap-Haïtien where the points are very close to the 45° line. In line with the scatter plot, the map depicting the ratio between predicted and observed values also contain values much closer to one.

This scaling exercise highlights the fact that the method could be valuable in estimating population densities in the absence of up-to-date census estimates in line with other methodologies such as Tatem (2017); Deville et al. (2014). The scaled population densities produced here could be used to understand population growth within the main metropolitan regions of the country. However in contrast with other methods, the methodology in this paper is dual in the sense that the labelling of meaningful locations required to avoid diluting the home location of an individual across multiple meaningful locations makes OD matrices a natural way of expressing and analysing the results.

Once evening population is estimated, the ratio between cell phone population and the total population is used to adjust each row of the OD matrix between evening and daytime clusters, thus producing an adjusted OD matrix. As such, it is assumed that all displacements from a given cell grid can be adjusted uniformly. However, it is possible to envisage situations in which such a condition does not hold. For instance, if a grid cell contains children and adults who have different destinations, it could be the case that child cell phone users might have a higher rate of adjustment than adult users as cell phone ownership among children is lower than for adults. Since this kind of analysis is not feasible with the present dataset, these kinds of biases are not considered in this paper.

5. Results

This section presents and explores population distributions obtained from CDRs using the methods described in the previous section and focusing on the metropolitan areas of Port-au-Prince and Cap-Haïtien. Since commuting behaviour is predicted for the whole country, an arbitrary rectangular surface given by a set of coordinates which might include a larger area than the administrative boundaries is used to define those metropolitan regions. The methods used in this report estimate stationary population distributions for two different categories. The first category is weekday daytime, which for most adults is likely to be work location. The second category is weekend and weekday evening, which for most people is likely to be home location. The labels day and evening time are used throughout this section to identify the first and the second category respectively.

It is important to keep in mind that the methods centre on stationary distribution of day and evening time populations, which is ideal for understanding regular commuting behaviour. The model does not take into account peak fluctuations which might be caused by particular events and days of the week nor does it attempt to assess seasonal fluctuations.

Second, the method might have an upward bias in the estimates of commuting distance since only those with commutes outside of their home tower coverage area will show up as having distinct home and work locations. On the other hand, this is mitigated by the fact that tower density is endogenous to phone user density, meaning that the algorithm will be more accurate where most of the phone users are located and by extension the population. In order to investigate this phenomenon we plotted the relationship between cell phone tower coverage area and percentage of commuters at night in Fig. 13. There is an inverted-U relationship between both variables. For towers with a small coverage area we see a lower percentage of commuters, which is likely caused by the fact that those towers are located in the centre of town where the share of commuters is lower since economic opportunities tend to be nearby. The percentage of commuters increases as the tower coverage increases since those towers are located further away from the centre of economic opportunities. However, for towers with high coverage areas, we find that the percentage of commuters start to fall likely caused by the upward bias present in the clustering algorithm. For those individuals who work and live within the boundaries of a cell phone tower, the clustering algorithm will identify a single cluster which conflates both home and work locations. Thus, only those users who commute long distances will have distinct clusters and for this reason the commuting distances will be upward biased in those towers.

In order to mitigate this bias, we calculated commuting distance metrics simple and weighted by the share of users in the cell phone tower that are deemed to commute. Those statistics are shown in Table 2. Overall, we find no significant differences whether the metrics are weighted or not. When looking at commuting distances stratified by distance from the centre of town we find that the weighted mean commuting distance is slightly higher than the simple mean for those commuters living between 1 and 5 km of the centre. We see an inverse relationship when looking at those living between 5 and 25 km of the centre. These trends are in line with the inverted-U relationship described above.

Port-au-Prince has a three-pointed star shape with a dominant centre surrounded by three substructures. The metropolitan area contains 3.5 million inhabitants, 42 percent of whom are considered commuters. Fig. 14 depicts the Port-au-Prince population distribution during day and evening time. The panel on the right shows population distribution during the evening time which most likely reflects home locations. The centre of Port-au-Prince sees the highest population densities, reaching up to 60,000 people per square kilometre during the evening. The density around the centre, which include neighbourhoods like Portail Léogane, Turgeau and Fort National, can be over 50,000 people per square kilometre reaching 55,000 people per square kilometre around Portail Léogane. Pétionville is the second most populated region in Port-au-Prince reaching densities of up to 50,000 people per square kilometre in its centre. To the west of the National Palace high population density is observed along Route Nationale 2 which leads to Carrefour. The

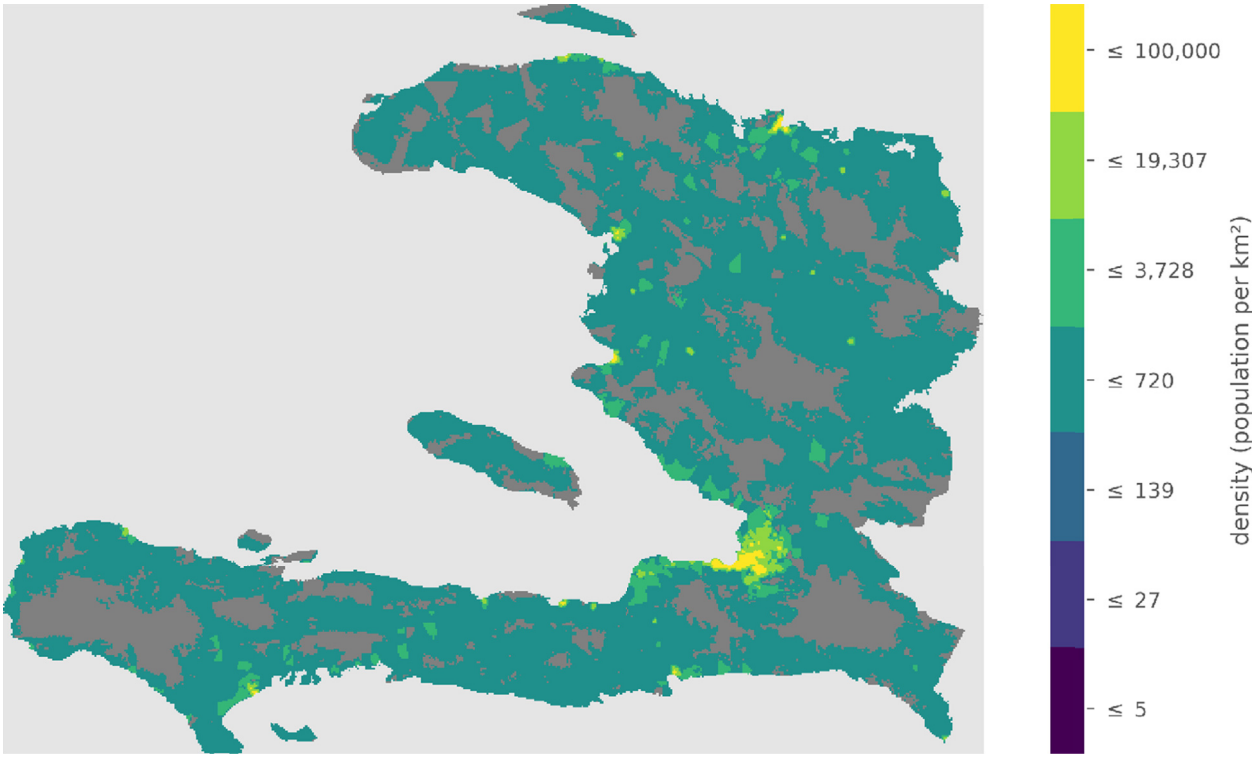


Fig. 11. Predicted population distribution — linear model.

Table 2
Key population and commuting statistics.

	Port-au-Prince			**Cap-Haïtien**		
Total Population	3.5 million			0.508 million		
Mean Trip	2.5 km			2.8 km		
Weighted Mean Trip	2.5 km			2.8 km		
Median Trip	1.1 km			1.6 km		
Weighted Median Trip	1.1 km			1.6 km		
	Non-Commuters	**Commuters**		**Non-Commuters**	**Commuters**	
		simple	weight.		simple	weight.
As percentage of total	58.14%	41.86%		60.51%	39.49%	
Mean Trip	–	4.5 km	4.5 km	–	4.7 km	4.7 km
Median Trip	–	3.1 km	3.1 km	–	3.3 km	3.3 km
Live less than 1 km from the centre	4.13%	2.86%		12.02%	6.13%	
Mean Trip	–	3.37 km	3.36 km	–	3.73 km	3.71 km
Median Trip	–	2.11 km	2.11 km	–	2.52 km	2.52 km
Live within 1 km and 5 km of centre	33.61%	31.67%		46.87%	56.81%	
Mean Trip	–	3.29 km	3.29 km	–	3.09 km	3.10 km
Median Trip	–	2.52 km	2.52 km	–	2.55 km	2.57 km
Live within 5 km and 25 km of centre	61.67%	65.27%		40.94%	36.98%	
Mean Trip	–	5.12 km	5.04 km	–	7.24 km	6.6 km
Median Trip	–	3.66 km	3.62 km	–	5.84 km	5.55 km

East side of Carrefour is the most populated part of the neighbourhood with densities lower than the centre of Pétionville. To the Northeast of the National Palace population is concentrated around Delmas. Past the airport, high relative densities are observed in Croix-des-Bouquets to the East and along Route Nationale 1 until the intersection with Route Nationale 3. Around this intersection one finds Cannan with around 10,000 to 15,000 people per square kilometre in its densest part. This is one of the most recent additions to Port-au-Prince, formed from temporary camps set up post-earthquake.

In terms of daytime population distribution one sees significant movement from the edges of Port-au-Prince towards the centre. The maps in Fig. 15 show the number of commuters as a percentage of day and evening time population. Not surprisingly, the centre of town sees commuters representing a substantial share of the population during daytime at around 72 percent but a very small share during the evening at around 40 percent. On the other hand, Carrefour and Canaan sees the opposite trend. It is interesting to note that along Route Nationale 1 and 8 which goes to Canaan and Croix-des-Bouquets respectively one sees a large increase in the share of commuters during the daytime.

The overall picture is one of concentration toward the city centre during daytime and inversely one of diffusion toward the outskirts during the evening. Fig. 16 shows that total population within 5 km of the city centre is about 5 percent higher during daytime than in the evening. If the focus is exclusively on commuters then the picture is even more striking with 46 percent of the commuters within 5 km

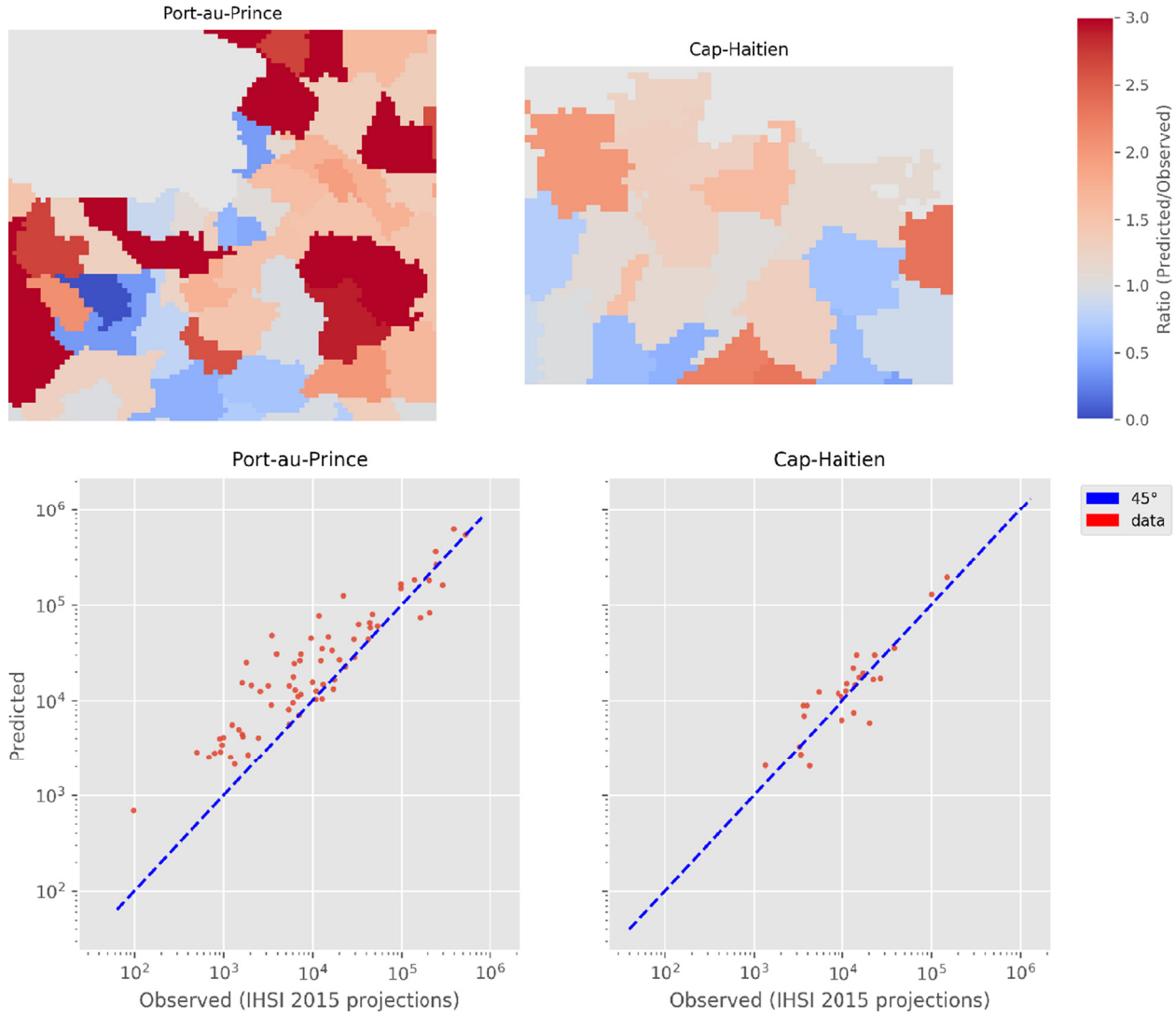


Fig. 12. Observed versus predicted population density at metropolitan level.

of the city centre during daytime versus less than 37 percent in the evening.

Using a GIS layer that contains neighbourhood classification derived from an automated building detection algorithm using remote sensing data (Antos et al., 2016), we compare in Fig. 17 population densities and commuting patterns across different neighborhoods. As expected, high population density areas align with high- built-up density zones, which have the second largest evening population density at around 18,000 inhabitants per square-kilometre followed by medium and low built-zones. Shanty zones have the highest evening population density at close to 25,000 inhabitants per square-kilometre. When looking at commuting, we find a major shift in population density in industrial zones which see an increase of 40 percent in population density during the day. High-density and other built-up zones also see population rises during the day. On the other hand, shanty- and medium-density built-up zones serve as dormitory areas. Some other areas with land cover classifications pointing at forests, bare soil, shrub and agriculture, are too small to provide illustrative comparisons. Those regions tend to be detected in the fringes of major neighborhoods. For this reason densities estimated in those areas are significantly higher than what would be expected. Although the GIS layer constitutes only an estimated classification, the agreement between this layer and the results presented in this paper brings strength to both pieces of evidence.

In the case of Cap-Haïtien, the centre exerts substantial pull. There are approximately 500,000 people living in Cap-Haïtien, 40 percent of whom are commuters. Fig. 18 shows that close to 60 percent of its population is concentrated within 5 km of the city centre. The attraction exerted by the city centre is powerfully illustrated by the fact that during the day 40 percent of all commuters can be found within 1 km of the city centre and nearly 80 percent within 5 km.

Despite the centre exerting substantial attraction over commuters in both metropolitan regions, most commuters will not commute a much greater distance than those living in the centre. In Port-au-Prince, it is estimated that the median trip is 1.1 km and 3.1 km if only commuters are considered. In Cap-Haïtien, the corresponding statistics are slightly higher at 1.6 km and 3.3 km respectively (see Table 2).

Those who live and work in the outskirts of Port-au-Prince, for instance, tend to travel smaller distances and they are likely spending their daytime in regions such as Croix-des-Bouquets and Pétionville. Commuting statistics vary across different neighborhoods in Port-au-Prince. In shanty zones commuters represent just below 8 percent of the total the evening population and they commute on average 3.4 km. On the other hand, commuters represent 26, 32 and 2 percent of the population in high-, medium- and low-density zones, commuting an average of 3.8, 4 and 4.8 km respectively.

In both urban areas median trip distances are low since it represent less than an hour of walking. On top of that, non-commuters comprise

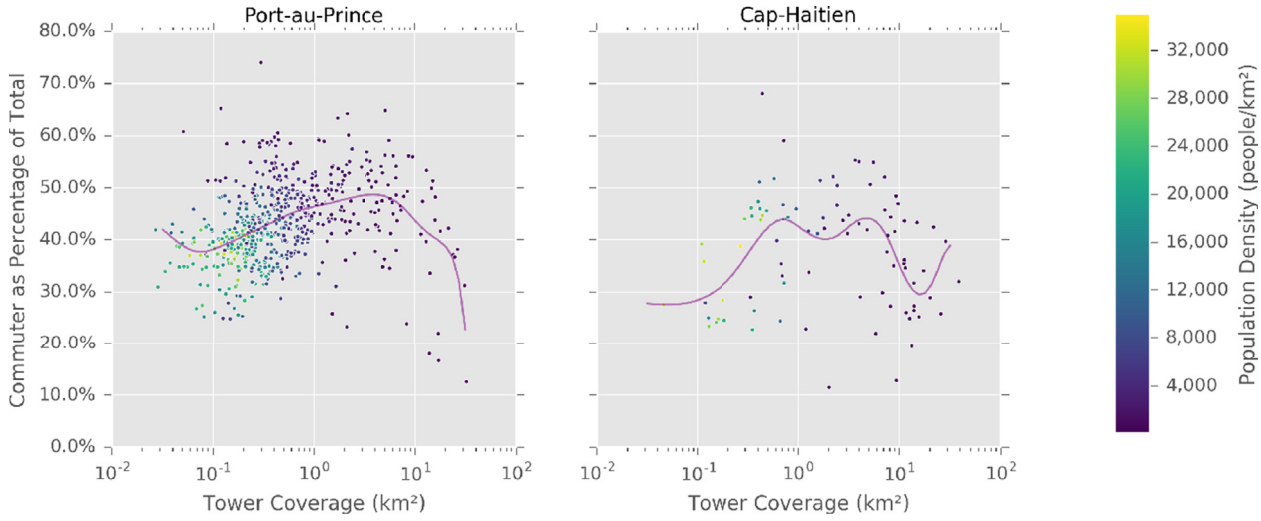


Fig. 13. Cell phone tower coverage bias.

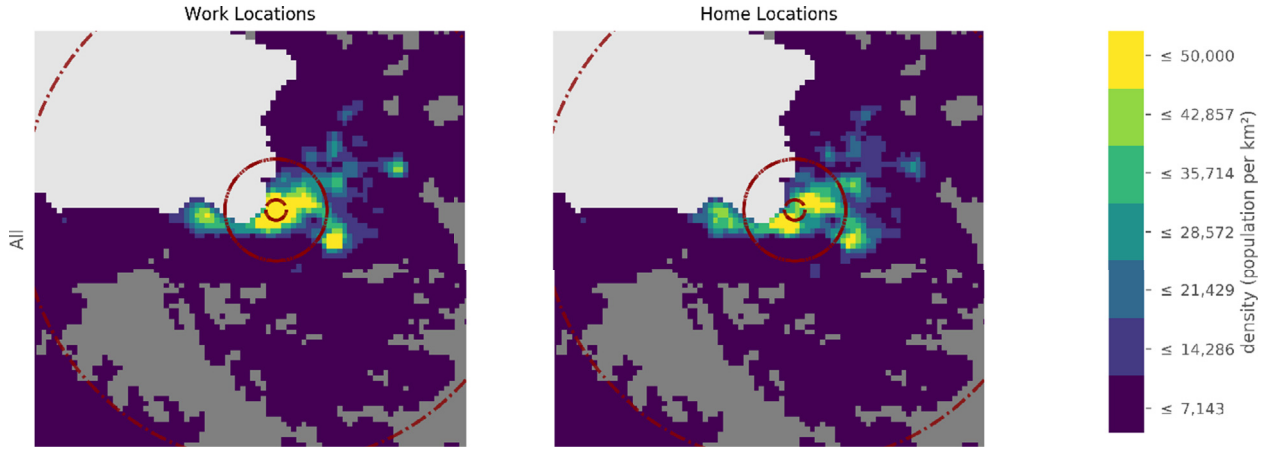


Fig. 14. Population distribution day versus evening and all versus commuters — Port-au-Prince. Concentric rings at 1, 5, and 25 km from city centre.

a significant share of the population reaching 40 percent in both cities. Given the characteristics of the algorithm, it is possible that for those individuals walking short distances to work, both home and work locations get conflated into a single estimate. As a consequence, the access to a large array of economic opportunities is likely low for the majority of the individuals. The fragmentary nature of these commuting patterns are indicative of fragmented labor markets which are unlikely to act as matchmakers, decreasing the probability of effectively pairing employers and employees.

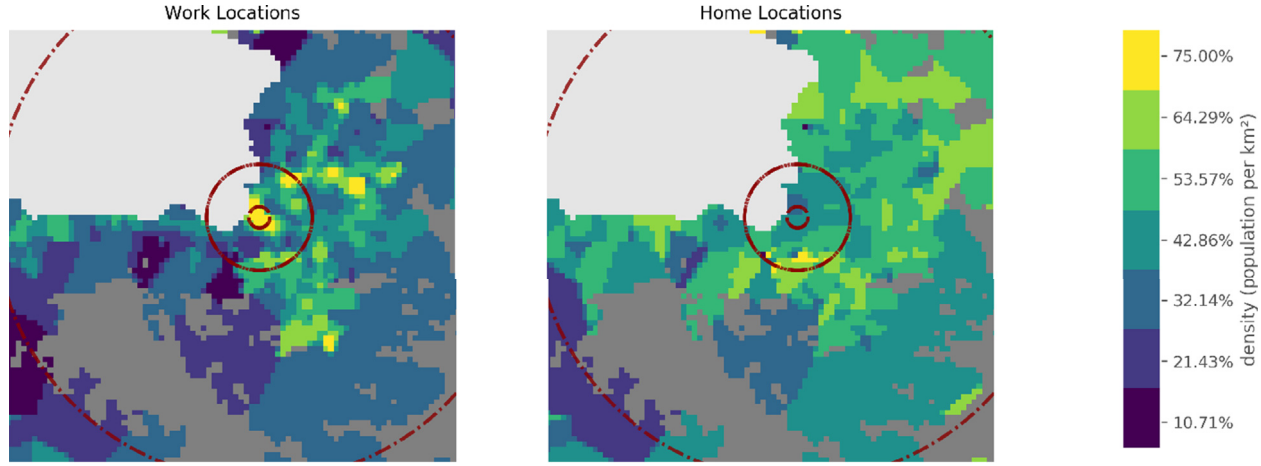


Fig. 15. Commuters as percentage of population — Port-au-Prince. Concentric rings at 1, 5 and 25 km from city centre.

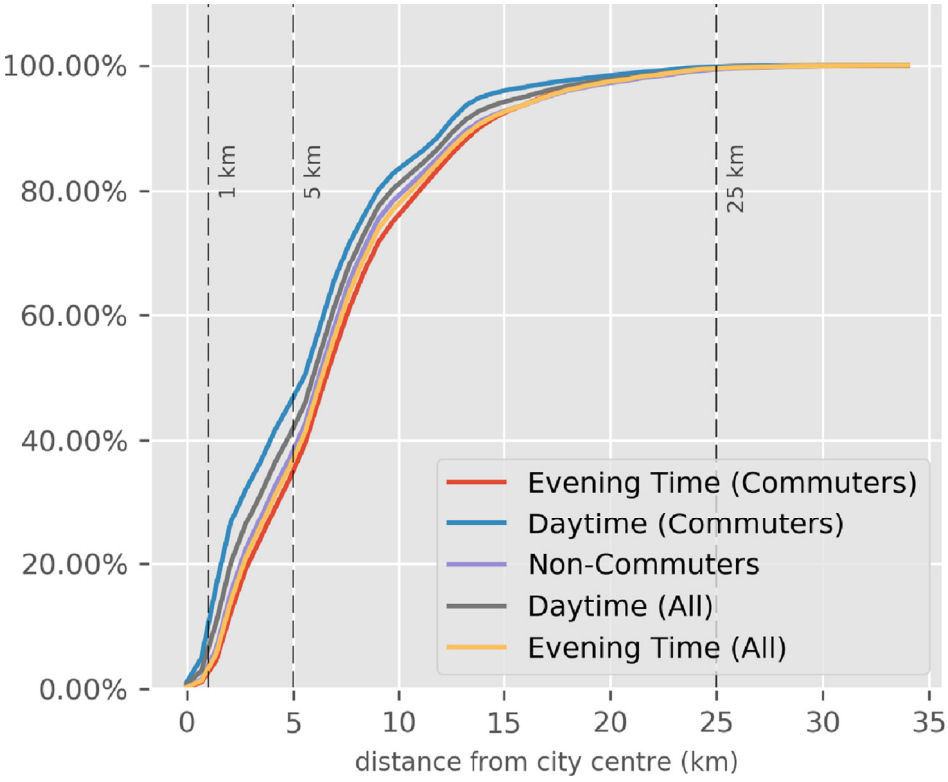


Fig. 16. Cumulative population distribution from city centre for different categories — Port-au-Prince.

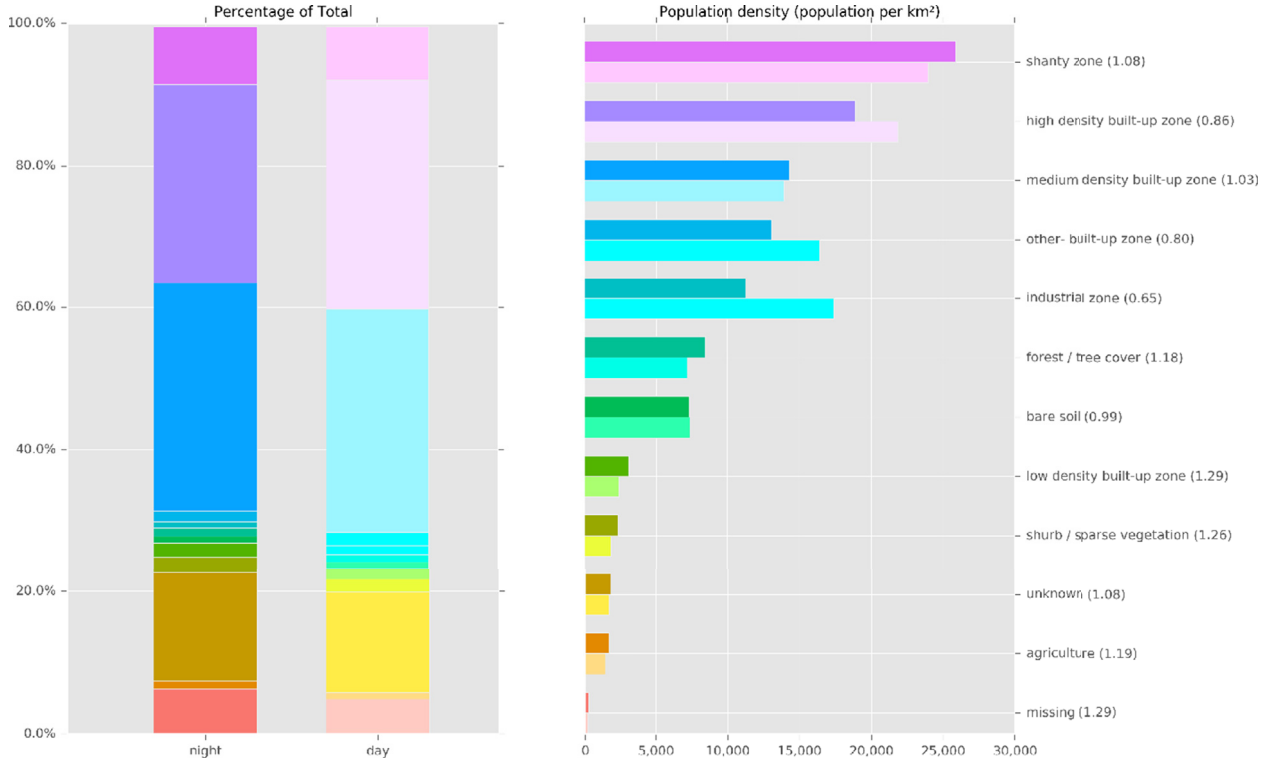


Fig. 17. Population distribution by neighbourhood type — Port-au-Prince.

This hypothesis is made stronger looking at the unequal distribution of travel distances. The distribution of travel distances for commuters that live the furthest from the city centre (beyond 5 km) in Port-au-Prince and Cap-Haitian is long-tailed, meaning that a number of commuters have the longest commutes as they are more isolated from economic opportunities and have to incur longer trips to reach these. The distribution of travel distances can be found in Figs. 6 and 11 in the appendix.

In line with our findings, Lozano-Gracia and Garcia Lozano (2017) reports that based on household expenditure surveys one can assume that 73.4 percent of the population in Port-au-Prince walk everywhere or do not travel since they report no expenditure on regular trans-

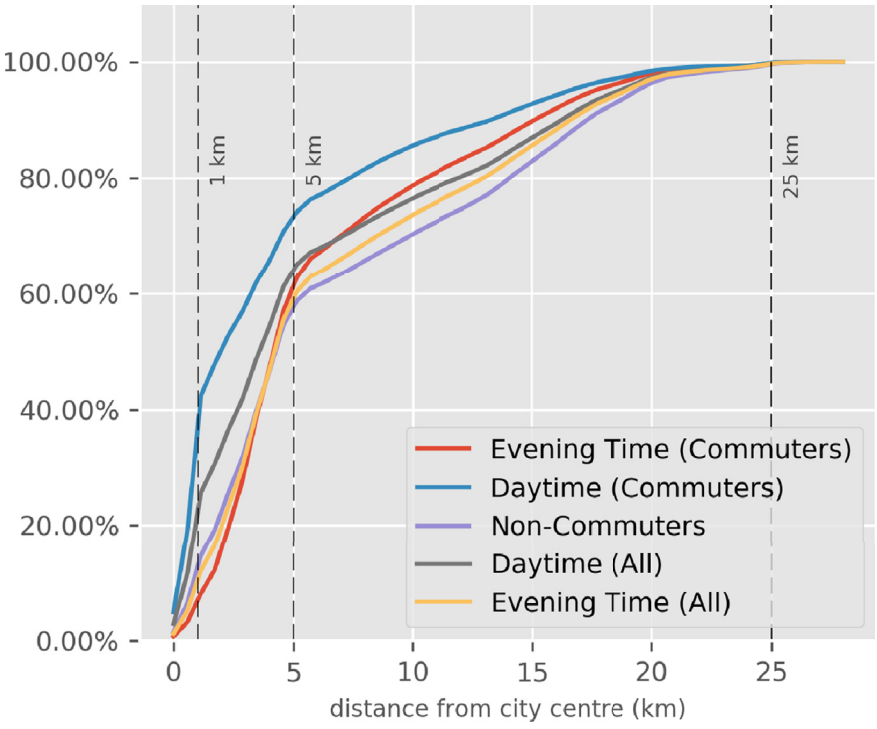


Fig. 18. Cumulative population distribution from city centre for different categories — Cap-Haitian.

port. Avner et al. (Feb/2017) finds similar dynamics in African cities which develop as collections of small, scattered neighborhoods with pockets of high-population density. The percentage of trips made by foot in Nairobi, Lagos, and Addis Ababa ranges from 30 to 45 percent, while reaching nearly 70 percent in Dar es Salaam and Kampala. Avner et al. (Feb/2017) also reports that average commuting distances is less than 6 km in Dar es Salaam, which is about double of the distance found for the cities analysed here. Lozano-Gracia and Garcia Lozano (2017) makes an extensive analysis of accessibility in Port-au-Prince and Cap-Haïtien using the OD matrices produced in this paper and comparing the results with other cities around the world.

A more detailed discussion of the results focusing on Port-au-Prince and Cap-Haïtien can be found in Appendix D.

6. Conclusion

This paper innovates from previous literature by developing a methodology for labelling meaningful locations in the absence of labelled data and by constructing an OD matrix depicting regular commuting patterns. Our work demonstrates the usefulness of using CDRs in data-poor environments. While such data cannot replace a well-designed travel survey, it provides enough information for a first assessment of the conditions on the ground and provides useful information to help inform policy and investments decisions. Future work should focus on assessing the precision of estimates such as those presented in this paper, as they compare to data collected from large and more detailed surveys, while also comparing the relative costs in terms of time and money.

The methodology developed in this paper could be validated with more traditional methods such as travel surveys, ideally by linking mobile phone and survey data. The efficiency and simplicity of the methodology developed in this paper could eventually allow for the development of real-time applications for monitoring commuting patterns in order to assist urban planners in their daily tasks.

The identification of meaningful places is also an input to an employment accessibility analysis and the identification of the degree of integration/fragmentation of labor markets in Port-au-Prince and Cap-Haïtien.

Acknowledgements

The work in this report has been made possible thanks to the financial contribution from three grants: support from the Global Facility for Disaster Reduction and Recovery (GFDRR) TF0A2693; grant from the World Bank's Jobs Umbrella Trust Fund, which is supported by the Department for International Development/UK AID, and the governments of Norway, Germany, and Austria, the Austrian Development Agency, and the Swedish Development Agency SIDA, TF0A2893; and grant from the Innovations in Big Data Analytics program, under the Global Data and Text Analytics Operations unit in the Global Themes Vice Presidency of the World Bank. The authors would also like to thanks the support provided by Digicel, the Comité Interministériel d'Aménagement du Territoire (CIAT) and the Centre National de l'Information Géo-Spatiale (CNIGS).

The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organisations, Flowminder, donors, data providers, or those of the Executive Directors of the World Bank or the governments/other institutions they represent.

Appendix A. Cell phone network and CDR

The CDR dataset used in this study originates from the normal operations of a mobile phone network and their intended purpose is as a billing instrument. A mobile telephone network is a geographically distributed radio network that enables communication via voice, text or data between two or more mobile devices — such as cell phones. Mobile devices communicate with each other through a network of base stations, which route the signal emitted from the originating mobile device to the destination mobile device. The position of a base station is fixed through most of its life, though some base stations can be relocated according to operational requirements. The base station is optimised according to a set of modulation parameters in order to reach an expected radio cell coverage.

Any base station can only handle a fixed amount of data per unit of time due to the limited amount of radio frequencies available. This phenomenon limits the amount of mobile devices that can establish communication with the rest of the network at any point in time. The network operator can adjust the expected base station coverage according to demand, such that more populated areas are serviced by a higher amount of low range base stations. As such, base station density tends to be positively correlated with population density. Base stations can be co-located. However, to avoid interference they might operate at different frequency bands and/or might cover different areas. A group of co-located base stations form a tower.

Due to environmental factors, the base station range is not fixed and tends to vary around its expected value. Its approximate coverage can only be discovered through field measurements and/or from simulations conducted as part of the radio planning and optimisation processes. A coarse estimation of the base station coverage area can be derived from antenna configuration parameters (e.g., antenna height, beam-width, tilt). When such parameters are not available, base station coverage can be approximated using a Voronoi tessellation which assigns for each base station location a polygon which contains all points in the plane for which such base station is the closest base station. Therefore, a Voronoi tessellation approach assumes that a cell phone would connect to the closest tower and that tower coverages do not overlap.

Mobile devices such as cell phones are uniquely identified via the International Mobile Subscriber Identifier (IMSI) which is usually provisioned in the Subscriber Identity Module (SIM card) of a mobile device. An IMSI is usually associated with a Mobile Station International Subscriber Directory Number (MSISDN) which is the standard phone number that users dial on their phone. The difference between the IMSI and the MSISDN is analogous to the difference between the IP address of a website which uniquely identifies a website on the internet and the website URL which is the mnemonic address used to access a website through the browser. A physical device such as a phone associated with the mobile device is uniquely identified via the International Mobile Station Equipment Identity (IMEI), the first 8-digit portion of the IMEI is known as the Type Allocation Code (TAC) which identifies the model and origin of the phone.

A mobile device spends most of its time in an “idle” state, i.e., not engaged in exchanging data with the network. The device switches to “active” when it is assigned radio resources to engage in data packets or signalling messages exchange. At any point in time only a small fraction of all mobile devices are found in the “active” state. Any “idle” mobile device is logically attached to a single network but is not assigned any radio resources.

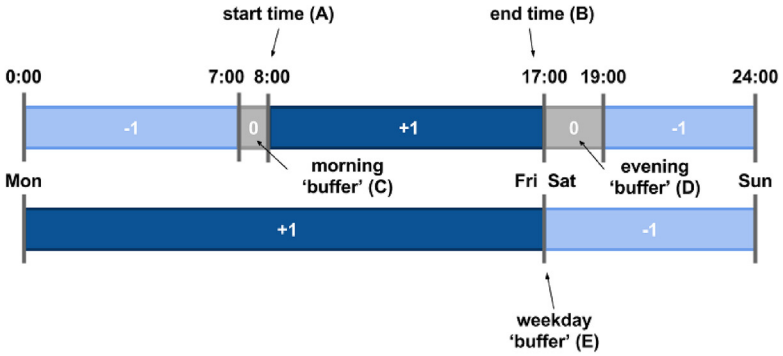
When “idle”, the mobile device makes decisions autonomously on which base station to listen to and when to switch to another base station when moving around. In this case, the mobile device is a passive receiver. Thus, the network has no way of detecting a base station change unless the mobile device decides to report this event explicitly. On the other hand, when the mobile device is “active” all decisions involving radio resources are taken by the network. As such, the network is able to track the position of active mobile devices at the base station level. The exact coordinates of a mobile device are never revealed to the operator and are, in fact, not required for the normal operations of the network.

In order to bill their customers, the network operator maintains a database of Call Detail Records (CDR). Every voice and data connection generates “tickets” that are sent to the billing system for charging purposes. This database is usually kept in a data warehouse for a long period of time which allows for revenue collection and dispute resolutions. CDR formats are not necessarily standardised across different networks, but the database will usually contain at least the MSISDN and IMSI for each SIM card, the starting time and duration of the connection, the type of connection and the base station ID of the starting base station where the connection was initiated.

Appendix B. Sensitivity analysis

There are a number of different parameters which control the labelling of clusters as day and evening time clusters including the start of the day, the end of the day, the buffer between the periods of the day and the buffer between weekends and weekdays. Each choice of parameter values results in a different way of labelling clusters as day and evening time clusters. We selected the parameters for the production of the final results according to sensible assumptions about the routines of the individuals in Port-au-Prince. Further, it was considered that Cap-Haïtien residents were likely to behave similarly.

The goal of the sensitivity analysis is to assess how robust the final results are to changes in key parameters. In order to carry out the sensitivity analysis, those parameters are varied within certain bounds and the resulting outputs are compared with a view to identifying systematic variations which could invalidate the final results. The results from the sensitivity analysis suggest that there is no systematic variation from the final choice of labelling parameters. Key parameters included those used to define the hour and the weekly score as follows:



Hour Score			
Description	**Legend**	**Unit**	**Values**
start time of the daytime period	A	hour	7:00, 8:00, 9:00
end time of the daytime period	B	hour	16:00, 17:00, 18:00
morning "buffer" duration	C	hours	2000, 1, 2
evening "buffer" duration	D	hours	2000, 2, 4
Weekday Score			
"buffer" duration between the end of the weekday and the start of weekend	E	days	0, 1

Fig. B.1. Sensitivity analysis parameters.

The variation in input parameters resulted in 162 different combinations of parameters. Each combination gave rise to a different neutral hour score, whereas the neutral weekday score was kept at zero. For each combination of parameters the clusters generated from the randomly selected sample of 10,000 users in Port-au-Prince were classified and gridded according to the methods outlined in the previous section.

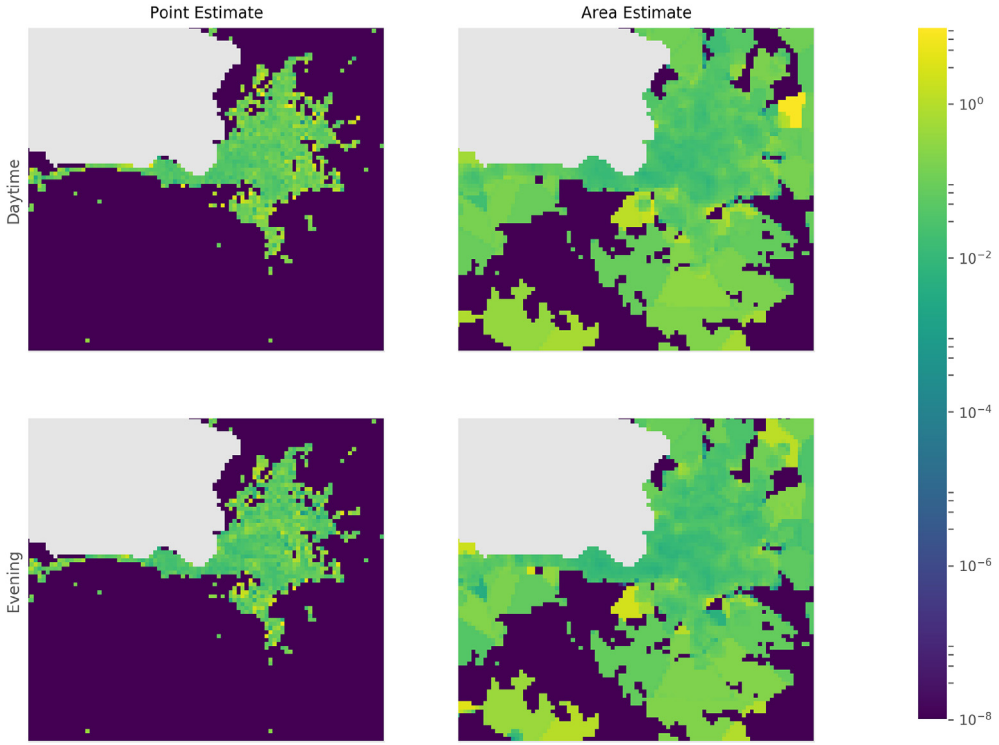


Fig. B.2. Distribution coefficient of variation (σ/μ) - log scale.

In order to understand whether the variation induced by different input parameters can be explained by any of the parameters described above, Fig. B.2 depicts the correlation between the parameters and the normalised grid cell area estimate values. Each grid cell is normalised by subtracting the minimum value of the grid cell and dividing by the grid cell range. The figure depicts in red and blue the values below and above the 80th

percentile of the coefficient of variation distribution, respectively. There does not seem to be any systematic variation across any input parameter variation both for day and night times.

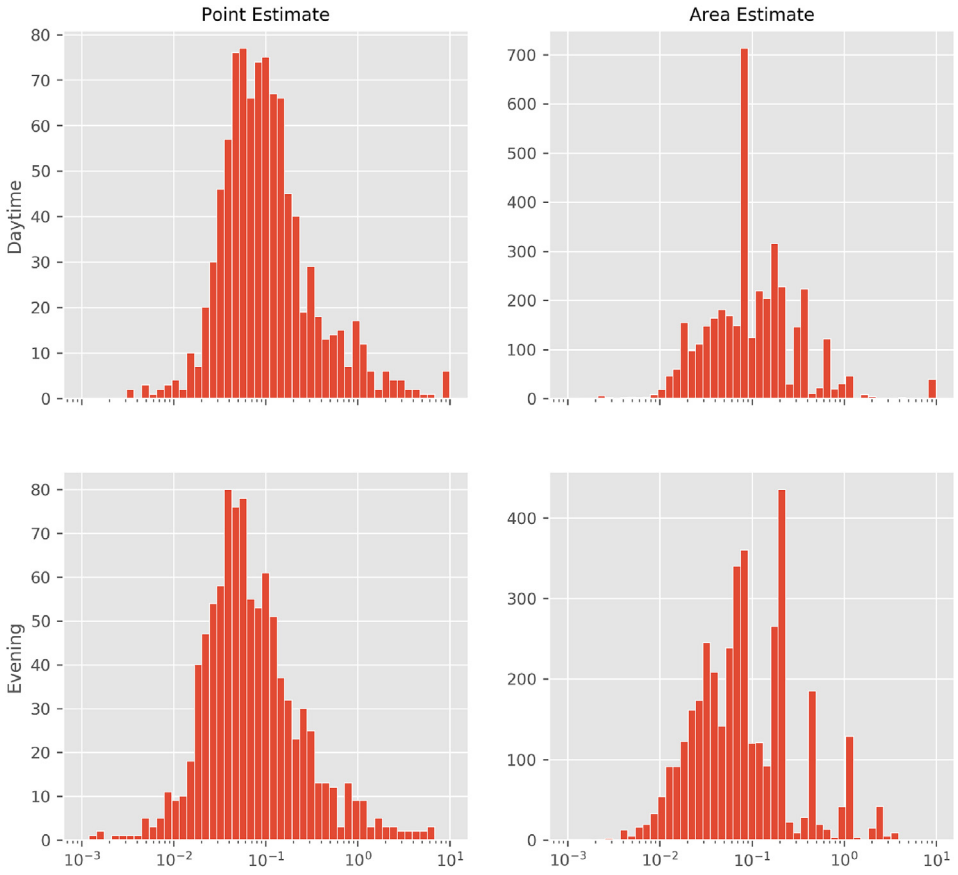


Fig. B.3. Coefficient of variation (σ/μ) — log scale.

Table Appendix B depicts the Ordinary Least Square (OLS) regression of the normalised grid cell estimates on the input parameters. Overall, the final results suggest that the input parameters have very little explanatory power on the final estimates. Even though any single coefficient of the regression might be significant, they are very close to zero suggesting that the variation induced by systematically changing the input parameters has no significant effect on the final results.

The conclusion of the sensitivity analysis is that although the day and evening time distribution of clusters are significantly different, systematic variation of the scoring criteria around its margin does not produce any systematic variation. That might be explained by the fact that the bulk of CDR events within day and evening clusters fall outside of the margins of the scoring criteria and moving the start and end of the day or adding buffers between them will not produce any systematic change in the final aggregated results. For this reason, the final input parameters as described in the previous section were chosen as the midway input parameters in the set of all the input parameter combinations.

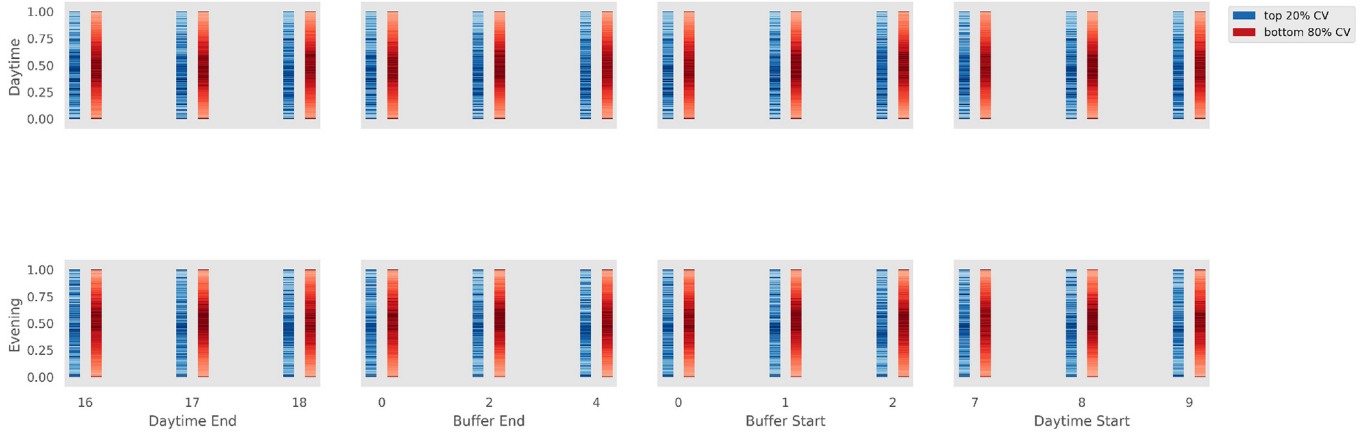


Fig. B.4. Coefficient of variation (σ/μ) - log scale.

Table B.1
OLS between normalised values and scoring parameters — significance at 0.05(*) and 0.001(**).

	Day time						Evening					
	Area			Point			Area			Point		
No. obs.	620,946			154,386			629,370			156,492		
R-squared	0.002			0.001			0.004			0		
Log-like.	−1.04E+05			−44,112			−1.03E+05			−47,483		
	Coef.	Std. error	z	Coef.	Std. error	z	Coef.	Std. error	z	Coef.	Std. error	z
Intercept	0.357 **	0.008	42.502	0.535 **	0.019	28.221	0.560 **	0.008	67.508	0.394 **	0.019	20.553
buffer_morning	−0.006 **	0	−12.81	0.005 **	0.001	5.426	0.016 **	0	37.236	−0.002	0.001	−1.465
buffer_night	0.005 **	0	23.08	0.001	0.001	1.039	−0.001 **	0	−6.521	0.001 *	0.001	2.44
buffer_week	0.015 **	0.001	20.515	0.007 **	0.002	4.076	−0.004 **	0.001	−5.785	0	0.002	−0.132
work_end	0.003 **	0	6.405	−0.002 *	0.001	−2.281	−0.001	0	−1.193	0.005 **	0.001	4.854
work_start	0.007 **	0	16.743	−0.005 **	0.001	−5.016	−0.014 **	0	−31.14	0	0.001	−0.418

Appendix C. Gridding

The different buffer sizes are compared in Fig. C.1 which shows cluster distribution during evening and day time for different buffer sizes of 0 m, 750 m and 1500 m for the randomly selected sample of 10,000 users in Port-au-Prince. As can be seen from these images, a small buffer size contains a lot of isolated points and stark variations, whereas the buffer size of 1500 m contains a lot of smoothed-out regions.

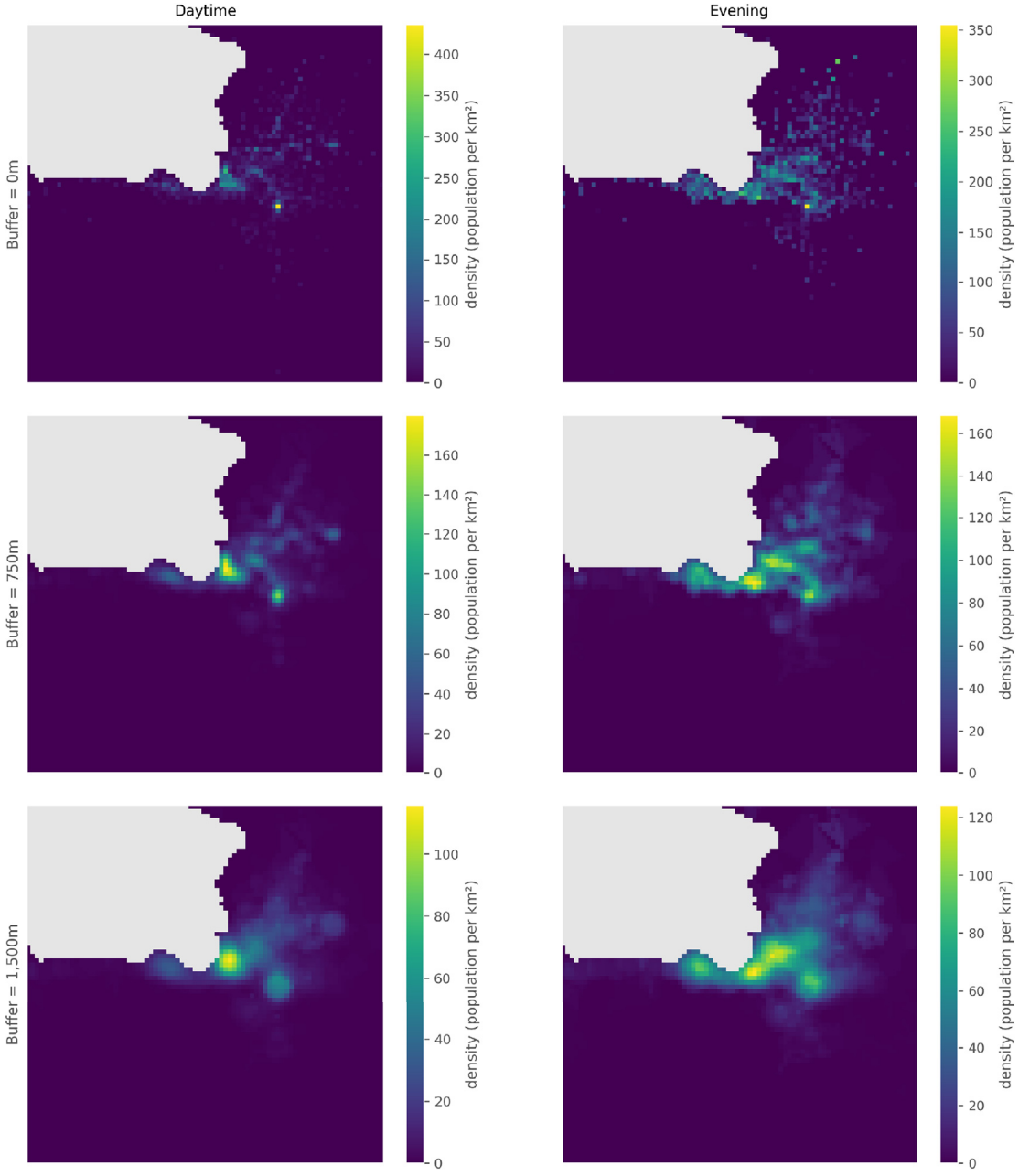


Fig. C.1. Daytime and evening cluster distribution — buffer variation.

Appendix D. Population distribution and commuting patterns

Appendix D.1. Port-au-Prince

Appendix D.1.1. Population distribution

Port-au-Prince contains 3.5 million inhabitants, 42 percent of whom are considered commuters. Fig. D.2 depicts the Port-au-Prince population distribution during day and evening time. The first row depicts all inhabitants of the metropolitan region and the second row depicts commuters only. The panel on the right shows population distribution during the evening time which most likely reflects home locations. The population of Port-au-Prince is scattered in a three-pointed star shape with its centre at the National Palace and the edges reaching to Carrefour on the West, to Pétionville on the south-east and to Canaan and Croix-des-Bouquets on the north-east. The centre of Port-au-Prince sees the highest population densities, reaching up to 60,000 people per square kilometre during the evening. The density around the centre, which include neighbourhoods like Portail Léogane, Turgeau and Fort National, can be over 50,000 people per square kilometre reaching 55,000 people per square kilometre around Portail Léogane.

Pétionville is the second most populated region in Port-au-Prince reaching densities of up to 50,000 people per square kilometre in its centre. To the west of the National Palace high population density is observed along Route Nationale 2 which leads to Carrefour. The East side of Carrefour is the most populated part of the neighbourhood with densities lower than the centre of Pétionville. To the North east of the National Palace population is concentrated around Delmas. Past the airport, high relative densities are observed in Croix-des-Bouquets to the East and along Route Nationale 1 until the intersection with Route Nationale 3. Around this intersection one finds Cannan with around 10,000 to 15,000 people per square kilometre in its densest part. This is one of the most recent additions to Port-au-Prince, formed from temporary camps set up post-earthquake.

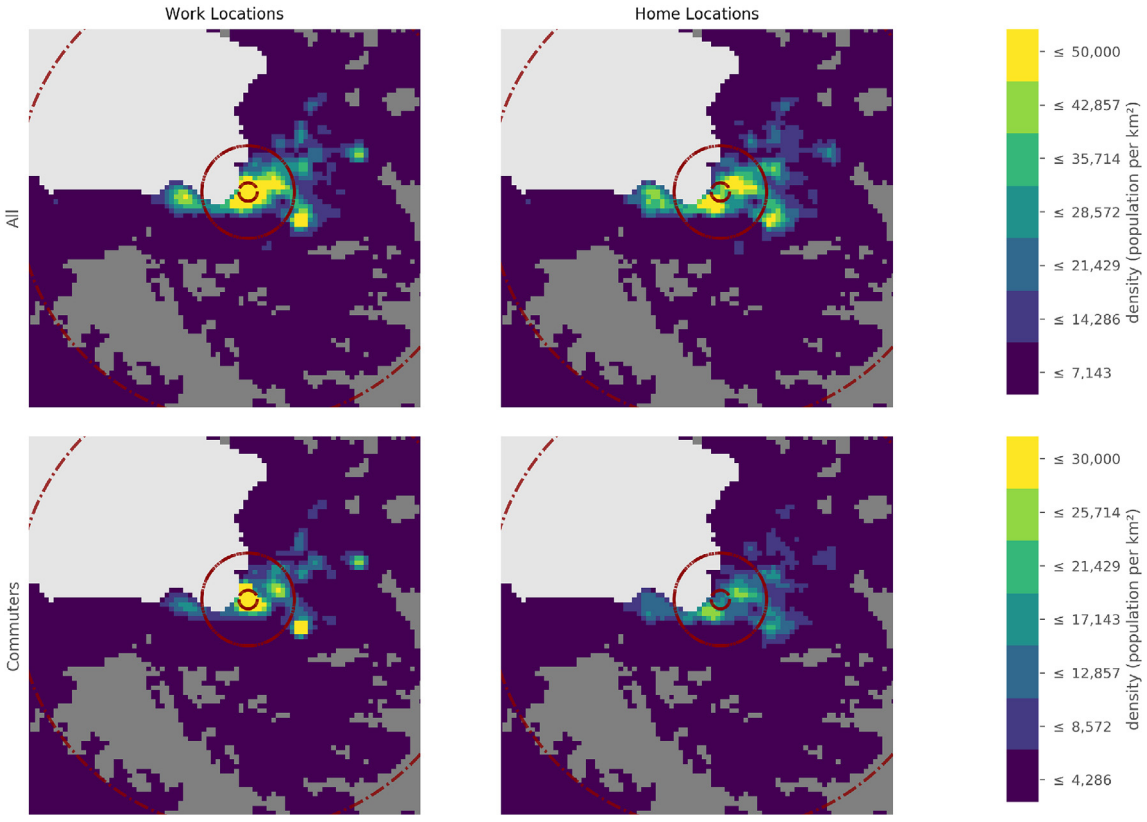


Fig. D.2. Population distribution day versus evening and all versus commuters — Port-au-Prince. Concentric rings at 1, 5, and 25 km from city centre.

In terms of daytime population distribution one sees significant movement from the edges of Port-au-Prince towards the centre. The movements are more clearly depicted in the second row of Fig. D.2, which depicts only commuters. Taking into account the whole population, the area around the National Palace can reach densities of up to 90,000 people per square kilometre during the day, which is 1.5 times higher than during evening time. Pétionville also sees some net increase during the day. Mean population density in the centre of Pétionville rises by 27 percent to just below 50,000 people per square kilometre from just below 40,000 during the evening. Likewise, some other regions to the north-east of the city centre also see some net increase, for example Saint Martin. Further north, the centre of Croix-des-Bouquets sees a net increase of 39 percent during daytime. On the other hand, some areas are notably residential. Carrefour and Canaan are two of those which see net decreases of 8 and 30 percent respectively.

Fig. D.3 can increase one's understanding of the main attracting and repelling regions. On the left and right-hand side the number of commuters as a percentage of day and evening time population, respectively, are depicted. Not surprisingly, the centre of town sees commuters representing a substantial share of the population during daytime at around 72 percent but a very small share during the evening at around 40 percent. On the other hand, Carrefour and Canaan sees the opposite trend. It is interesting to note that along Route Nationale 1 and 8 which goes to Canaan and Croix-des-Bouquets respectively one sees a large increase in the share of commuters during the daytime.

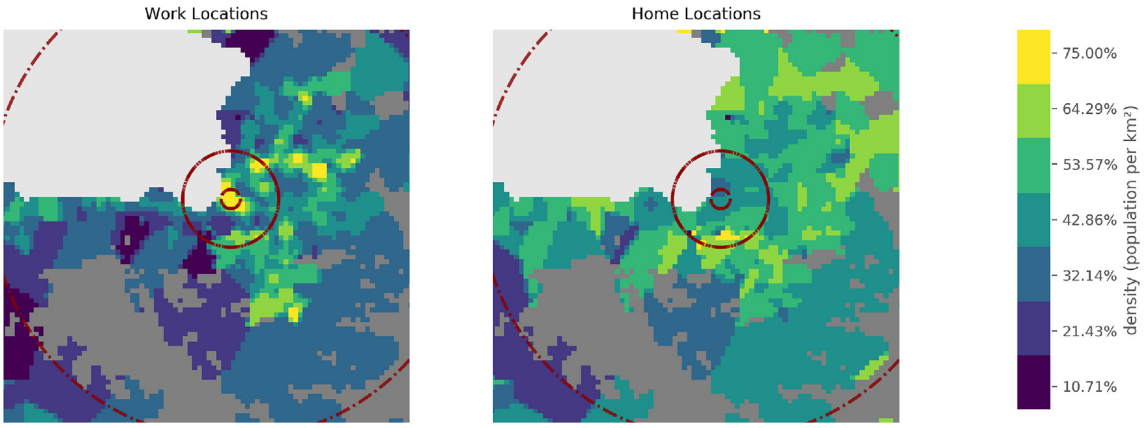


Fig. D.3. Commuters as percentage of population — Port-au-Prince. Concentric rings at 1, 5 and 25 km from city centre.

Turning one's attention to population distribution in relation to the city centre, Fig. 16 can be helpful. The graph depicts the cumulative percentage of inhabitants who are located at a given distance from the centre of Port-au-Prince for different categories. The image suggests that for every kilometre from the city centre one sees approximately an extra 8 percent of the total population of Port-au-Prince. There are no major inflection points until one reaches 10 km from the city centre, when the cumulative distribution begins to tapers off. The lack of any other major breakpoint is the result of the population of Port-au-Prince being distributed along a three-pointed star with three smaller focal points in Carrefour, Pétionville and the region comprised of Delmas, Croix-des-Bouquets and Cannan.

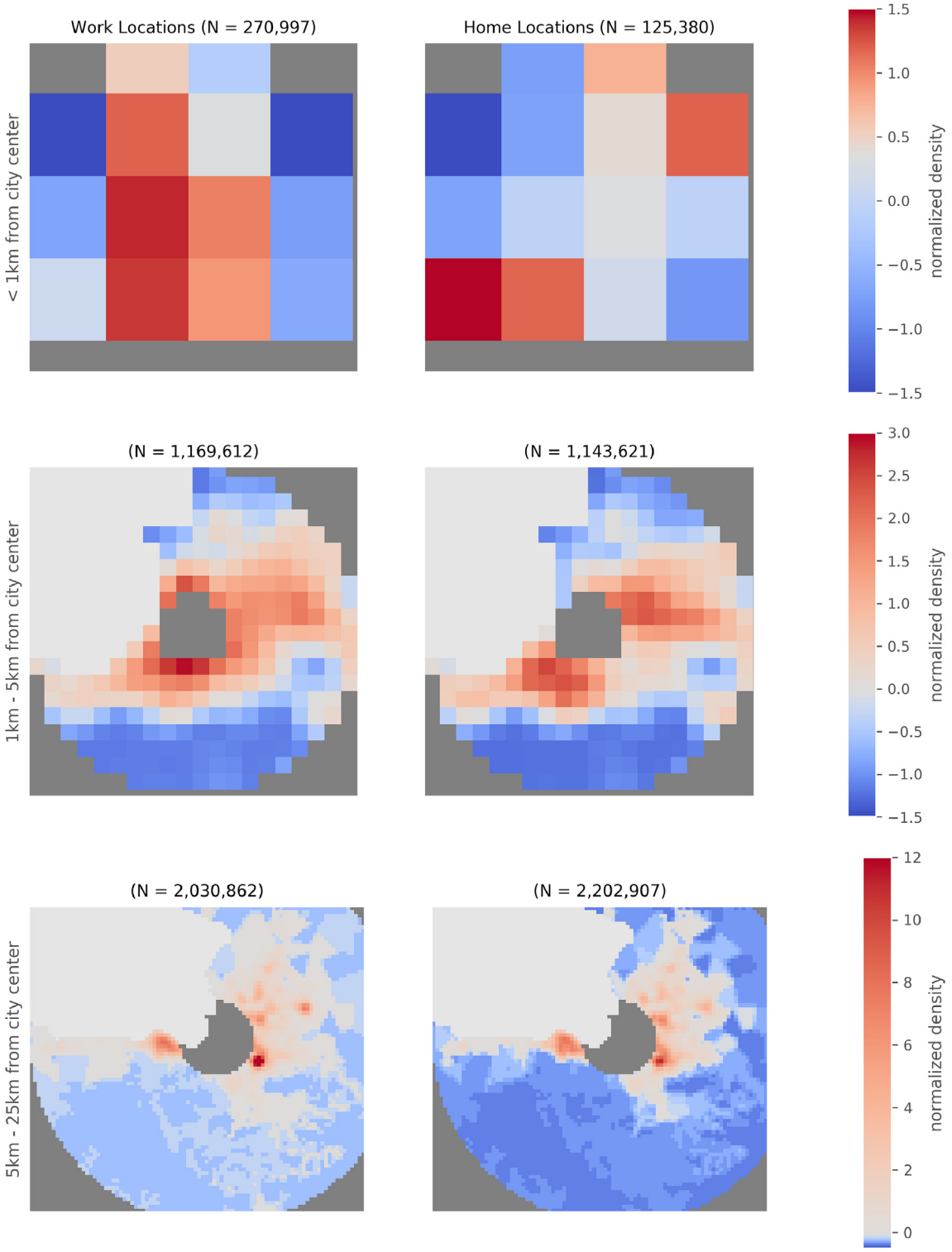


Fig. D.4. Relative population distribution within buffers (1, 5 and 25 km from city centre) — Port-au-Prince. Total population in brackets.

Nevertheless, one notices that there is a clear move towards the city centre during daytime. At any given distance from the city centre total population is always smaller during evening time than during daytime and this contrast is even more marked for data on commuters only.

Fig. D.4, depicts the relative population distribution of Port-au-Prince along concentric circles centred in the city centre at the National Palace with radius of 1, 5 and 25 km. Each map depicts the normalised density calculated over the density of each buffer in order to highlight the main inhabited areas in those buffers during the day and night.

The second row of Fig. D.4 depicts the relative distribution of the population located at least one kilometre and at most five kilometres from the National Palace. It can be seen that during daytime, population is concentrated around the town centre, but during the evening time the population disperses towards two circles concentrated around Portail Léogane and another around Fort National.

The last row of Fig. D.4 depicts the third buffer which is located at least 5 and at most 25 km from the city centre. In agreement with the hypothesis put forward above, population seems to be concentrated around three secondary centres namely Carrefour, Pétionville and the north of Port-au-Prince. Despite being a residential region, Carrefour still hosts a number of people during the daytime, since only about 40 percent of the population seem to commute from this area during the day. As discussed above, Pétionville hosts a significant number of people during daytime due to a large influx of commuters. During the evening, the population is slightly more dispersed from its centre. Finally, the region north of Port-au-Prince contains pockets of people around Delmas, Croix-des-Bouquets and Cannan. Similarly to Pétionville, the centre of Croix-des-Bouquets sees a large relative increase in people during the day.

Appendix D.1.2. Population flows

It is worth analysing the flow of people from each of the buffers in order to understand usual commuting behaviour. Fig. D.5 depicts the flow of commuters from each specified buffer. The column on the left depicts the daytime (work) location of the individuals who live in the selected buffer. The column on the left depicts the evening time (home) location of the individuals who work in the selected buffer. For instance, the column on the right shows where those people living in the centre of Port-au-Prince go to work, whereas the column on the left shows where those people working in the centre of Port-au-Prince live.

The first row of Fig. D.5 depicts commuters from the centre of town. It is possible to see that most commuters who live in this area travel to Delmas and Pétionville during the day. Fig. D.3 in the previous section shows that the share of commuters in the centre of Port-au-Prince is rather small, which is likely due to the fact that most people who live in the centre of Port-au-Prince also work there. The centre of Port-au-Prince sees a huge influx of people from different areas during the day including Carrefour and Martissant on the West, Pétionville on the Southeast and Delmas on the Northeast.

Fig. D.6, depicts the distribution of distance travelled for each buffer. The column on the left depicts the distance travelled by individuals who live in the selected buffer, thus showing how far they likely travel from home to their workplace. The column on the right depicts the distance travelled by individuals who work in the selected buffer, thus showing how far they likely travel from their workplace to home. The blue legend depicts those who travel outside of their buffer and the red one those who remain inside. The graph also shows the amount of people undertaking each commute.

The first row of Fig. D.6 reaffirms the claim that many more people travel into the city centre than to the outside of it. There are about 4.5 commuters who go into the city centre during the day for every commuter who goes out of the city centre. It can be seen from Fig. D.6 that most commuters to the city centre will travel a distance between zero and five kilometres. There is a relative high number of people who will travel between five and ten kilometres, and those people are likely located in the edges of the three-pointed star discussed previously.

The second row of Fig. D.5 reiterates that most people within a buffer of at least one and at most five kilometres from the city centre will commute to the city centre to work and a significant share will also commute to Pétionville. There is a considerable number of people who work and live in the same buffer, as depicted in the left pane of Fig. D.6. The same figure also shows that most commuters will not commute a much greater distance than those living in the centre. In fact, the trend seems to go the other way. This buffer is a net importer of people during daytime, with about a five percent more people coming in during the day than leaving. People working in this buffer will travel from two to five kilometres to work and seem to flow from the same directions as those coming to work in the centre of town.

People who live in the most distant buffer see similar movements of large number of people going to work in the centre of Port-au-Prince and Pétionville. The region has almost 20 percent more commuters during commuting to work in the morning than commuting from work during the evening. Nevertheless there is more variation, with people living in this buffer going to work as far as Croix-des-Bouquets and Cannan.

The histogram shows that there is a number of people who travel relatively greater distances to work. The graph shows that those who tend to travel longer distances to work also tend to go out of the buffer during the day, suggesting that those are the ones who travel to the centre of town. A large proportion of those people likely travel from Carrefour. On the other hand, those who live and work in this buffer tend to travel smaller distances and they are likely spending their daytime in regions such as Croix-des-Bouquets and Pétionville. The pane on the left of Fig. D.5 confirms this, as the many well lit places all over the buffer indicate that people who work in this buffer also tend to live in the same region and travel smaller distances.

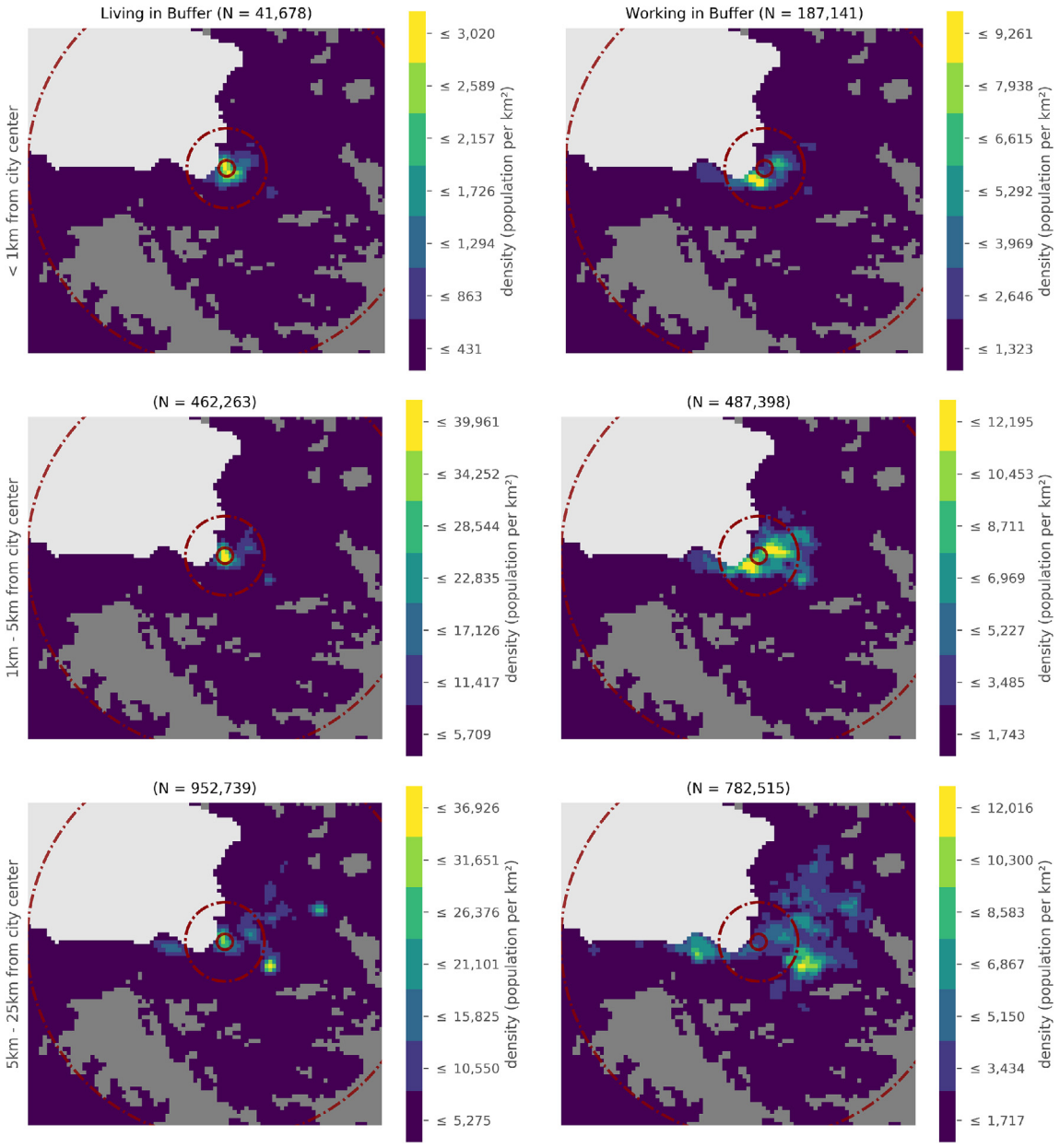


Fig. D.5. Flows from buffer (commuters only) day versus evening time — Port-au-Prince. Concentric rings at 1, 5 and 25 km from city centre.

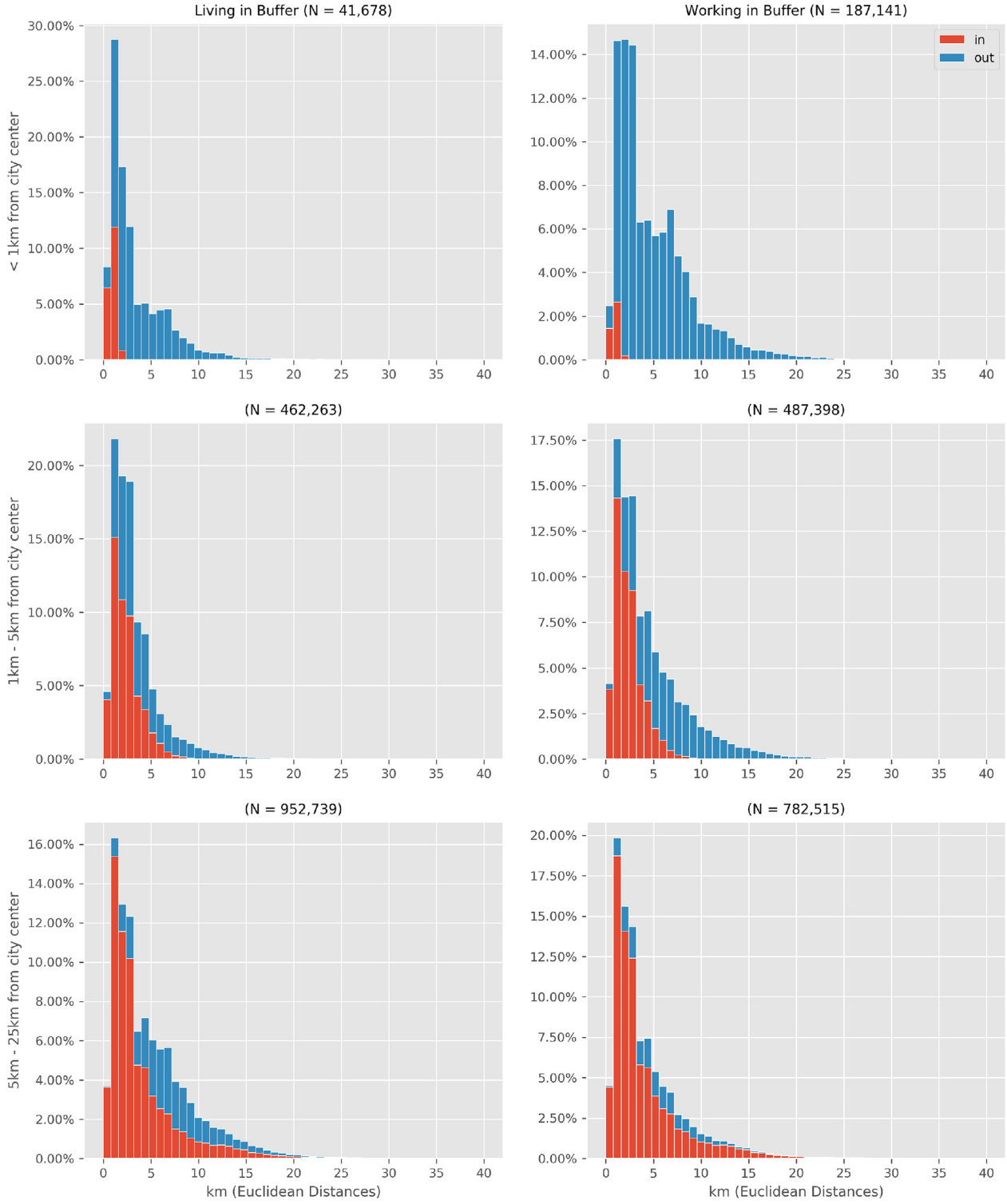


Fig. D.6. Distribution of euclidean distances travelled from buffer (commuters only) day versus evening time — Port-au-Prince. Concentric rings at 1, 5 and 25 km from city centre.

Appendix D.2. Cap-Haïtien

Appendix D.2.1. Population distribution

There are approximately 500,000 people living in Cap-Haïtien, 40 percent of whom are commuters. The vast majority of the population lives towards the centre of town on the West side of the bay and Mapou River and along the South bay east of Mapou River. The highest evening population density can be found around La Fossette, a popular middle to low income residential neighbourhood to the west of River Mapou, and close to the only bridge linking both sides of town. Densities in these regions can reach just below 40,000 people per square kilometre. On the other side of the bay, density is higher around Petite Anse, where it oscillates around 20,000 people per square kilometre.

Cap-Haïtien is not as big as Port-au-Prince. Population density decreases as one travels south along Mapou river. Population density is high within a narrow two kilometre strip west of Mapou river and six kilometres from the mouth of the river until Haut-du-Cap. Outside of Cap-Haïtien,

population is relative higher along Route Nationale 1 from Vaudreuil to Moustique to the South-West, in Quartier Morin, Limonade and Trou-du-Nord to the South-East and in Milot to the South. Arguably most of those regions are not part of the metropolitan area of Cap-Haïtien but are nevertheless mentioned here since they are covered by the grid under consideration. Population density in those areas fluctuates between 500 and a 1000 people per square kilometre and is particularly higher in Vaudreuil and Trou-du-Nord. Outside of those satellite regions, density drops dramatically reaching below 500 people per square kilometre.

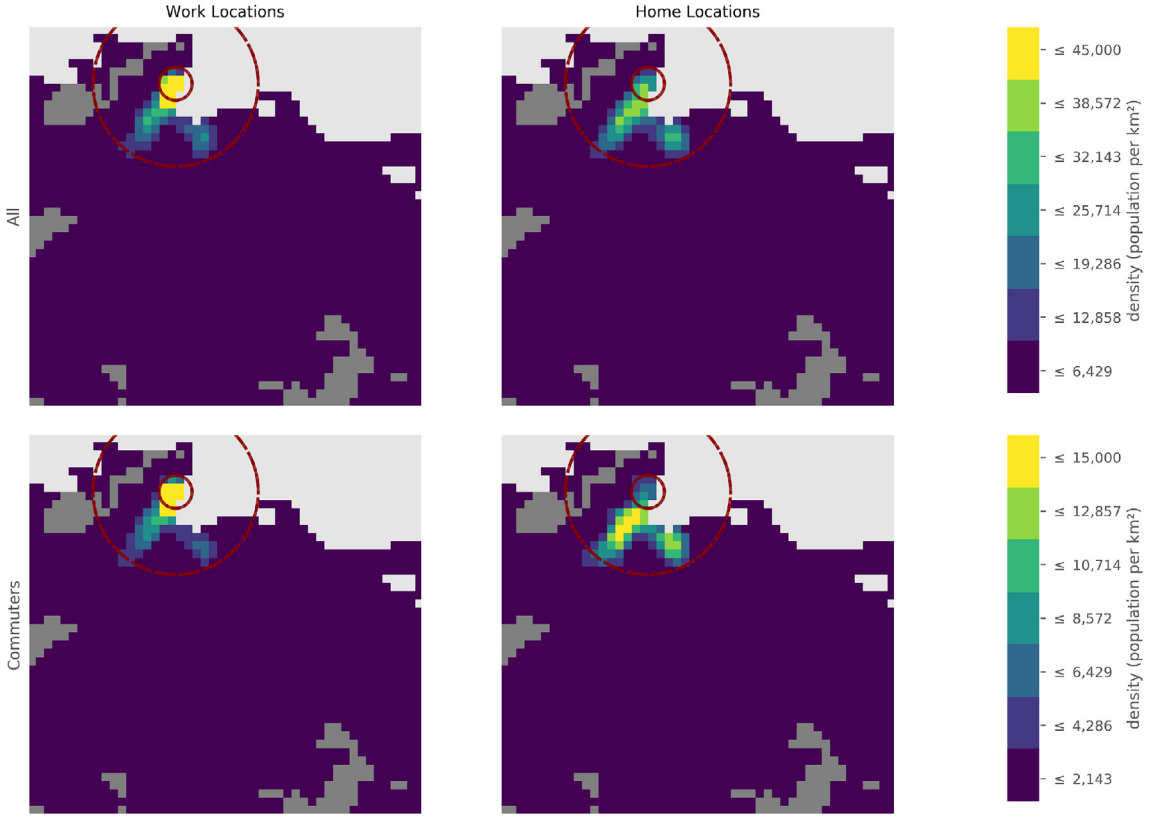


Fig. D.7. Population distribution day versus evening and all versus commuters — Cap-Haïtien. Concentric rings at 1, 5, and 25 km from city centre.

During daytime most commuters in Cap-Haïtien head towards the business district right in the centre of town where population density can be as high as 80,000 people per square kilometre similar to Port-au-Prince. Petite Anse sees significantly fewer commuters than the centre of Cap-Haïtien and average population density actually drops during daytime by about 24 percent from 13,000 people per square kilometre during the evening to about 10,000 during the day. Driving south from the business district along River Mapou, daytime density gradually decreases but at a faster rate than during the evening.

The centre of Cap-Haïtien sees a substantial increase in the relative number of commuters during the day as can be gleaned from Fig. D.8. During the day, commuters represent just below 70 percent of the total population in the centre, dropping to just a quarter of the population during the evening. The region of Petite Anse sees the opposite trend with the share of commuters dropping to below 30 percent during the day and rising to above 40 percent during the evening.

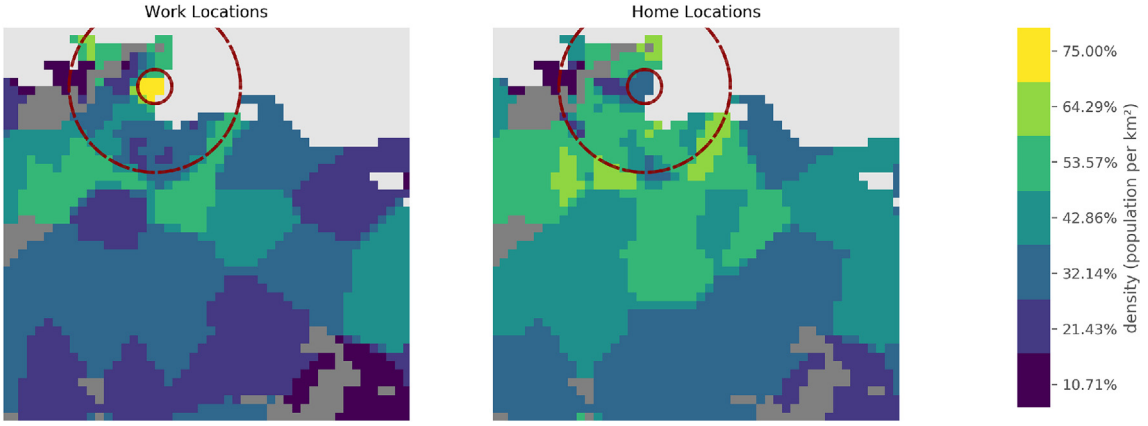


Fig. D.8. Commuters as percentage of population — Cap-Haïtien. Concentric rings at 1, 5 and 25 km from city centre.

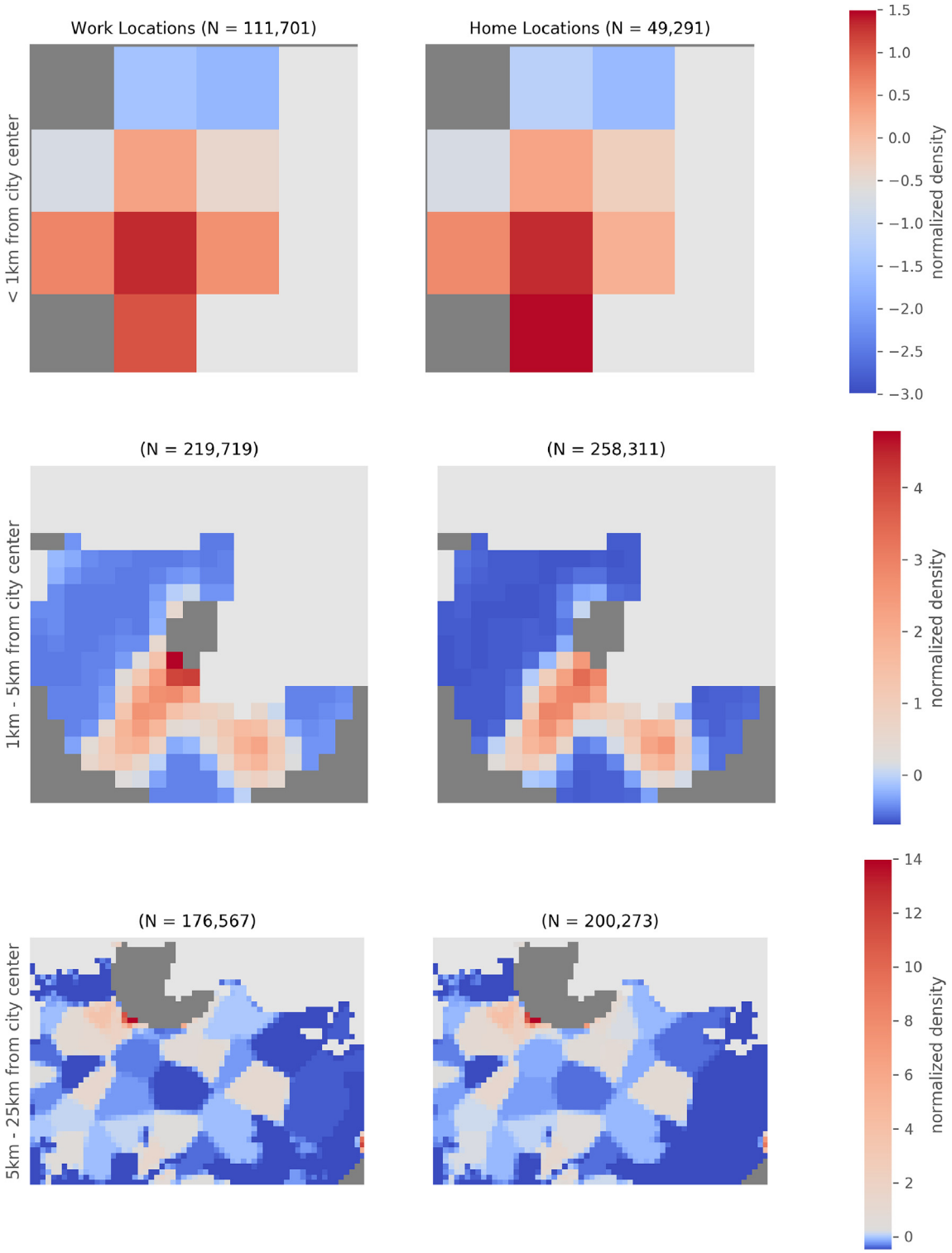


Fig. D.9. Relative population distribution within buffers (1, 5 and 25 km from city centre) — Cap-Haïtien. Total population in brackets.

As previously mentioned, Cap-Haïtien is characterised by a vast concentration of people along the thin strip running two kilometres west of River Mapou and five kilometres south from the mouth of River Mapou up to Haut-du-Cap. This is reflected in Fig. 18 depicting cumulative population distribution from the city centre with an inflection point at five kilometres.

The figure suggests that about 60 percent of the population in Cap-Haïtien can be found within five kilometres of the city centre at a rate of about 12 percent of the total population for every kilometre from the centre. When considering the distribution of commuters during the day, one finds an even greater concentration of people towards the centre of town. During the day, 40 percent of all commuters can be found within one kilometre of the city centre and about 80 percent within 5 km.

The relative population distributions within 1, 5, and 25 km buffers are depicted in Fig. D.9. In line with the analysis above, most of the population living and working within one kilometre of the city centre is located towards the south of the business district in La Fossette and closer to the only bridge linking the two sides of town. During the daytime, the centre of Cap-Haïtien sees its population almost double.

Likewise, between one and five kilometres from the city centre, during daytime, most of the people are located close to the only bridge linking both sides of Cap-Haïtien. The buffer sees a net loss of people during the day, most of whom likely go to the centre of town. During the evening, the population disperses towards the south of River Mapou and towards Petite Anse. This buffer holds the largest number of people during the night out of all the other buffers considered.

At distances bigger than 5 and less than 25 km from the city centre, one finds population scattered across small villages in what is mostly rural areas. Most of the population can be found along Route Nationale 1 with particular focus on Vaudreuil which is just south of Haut-du-Cap along this motorway. Other focal points include south of Petite Anse in Quartier-Morin, Limonade and the small village of Trou du Nord in the Southeast. This buffer holds during the night about 20 percent fewer people than the previous buffer but loses approximately 40 percent less people during the day.

Appendix D.2.2. Population flows

The flow of commuters from each buffer during the day and evening time is depicted in Fig. D.10. The distribution of distances travelled is depicted in the following Fig. D.10. The first row of Fig. D.10 left pane shows that a number of commuters who live in the centre of town tend to go to Petite Anse during the day and some others tend to travel further south along River Matou. The centre of Cap-Haïtien sees about six people commuting into the area during the day for every commuter who travels outside of it.

The histogram (first row, right column) shows that about 70 percent of the trips to the centre are less than five kilometres. Most of those trips are from commuters who live in the second buffer either further south along River Matou or in Petite Anse. The second row left pane of Fig. D.10 shows that the vast majority of commuters in the second buffer takes that direction. About 95 percent of the commuters (Fig. D.10, second row, right column) in the second buffer travel less than 5 km to work.

The histograms in the first and second row on the right hand side show two peaks. The second peak on the right suggests that there is a small fraction of people who travel a considerable distance during the day of around 15 km. The histogram in the third row of Fig. D.11 depicting travelled distance in the morning from the third buffer confirms that indeed there is a bimodal distribution with two peaks at 5 and 15 km.

This fact suggests that there are two groups of people who travel to the centre of Cap-Haïtien. The first one includes the majority of commuters who live concentrated around Vaudreuil and south of Petite-Anse. The second group is smaller and is scattered around the countryside in small villages such as Quartier-Morin, Milot and Trou-du-Nord in and around the main access routes to Cap-Haïtien including Route Nationale 1, 3 and 6. On the other hand, those who work in the third buffer tend to travel around five kilometres back home with a significant number of them working in Vaudreuil and south-east of Petite Anse (Quartier-Morin, Limonade and Trou-du-Nord) and returning back home to the second buffer.

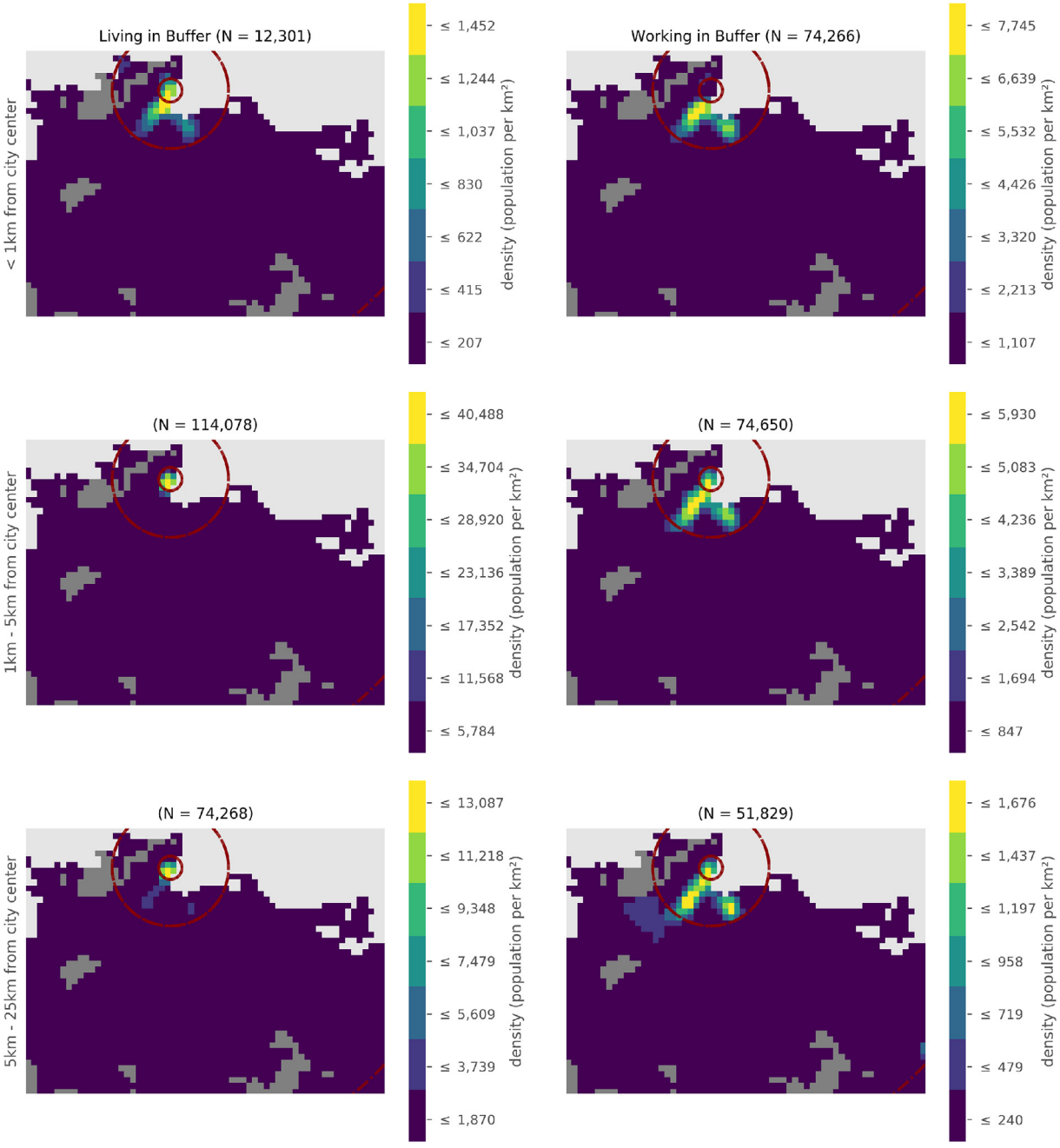


Fig. D.10. Flows from buffer (commuters only) day versus evening time — Cap-Haïtien. Concentric rings at 1, 5 and 25 km from city centre.

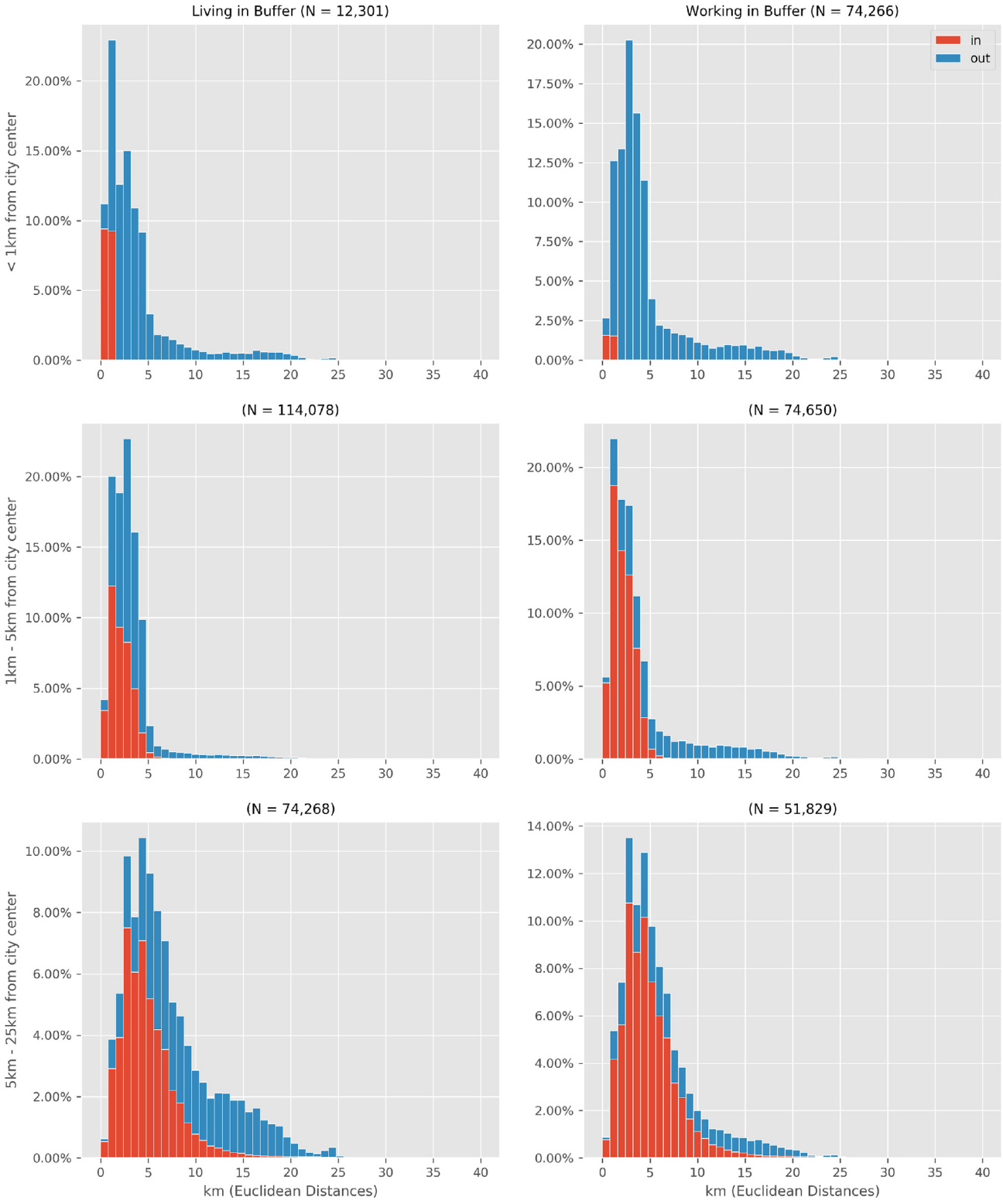


Fig. D.11. Distribution of euclidean distances travelled from buffer (commuters only) day versus evening time — Cap-Haïtien. Concentric rings at 1, 5 and 25 km from city centre.

Appendix E. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.deveng.2018.03.002.

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., Apr. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. ResearchGate 17 (1), 3–27.

Antos, S.E., Lall, S.V., Lozano Gracia, N., Dec. 2016. The Morphology of African Cities. Tech. Rep. WPS7911. The World Bank.

Avner, P., Henderson, J.V., Lall, S.V., Baruah, N., Bernard, L., Bird, J., D'Aoust, O., Lall, S.V., Feb/2017. Disconnected land, people and jobs. In: Africa's Cities: Opening Doors to the World. The World Bank, pp. 62–84.

Calabrese, F., Ratti, C., Di Lorenzo, G., Liu, L., 2011. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. 10, 36–44.

CONATEL, 2016. Tableau de bord du secteur de la téléphonie mobile. Tech. rep.. Conseil National des Télécommunications.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., Nov. 2014. Dynamic population mapping using mobile phone data. PNAS 111 (45), 15888–15893.

González, M.C., Hidalgo, C.A., Barabási, A.-L., Jun. 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.

Graells-Garrido, E., Saez-Trumper, D., Feb. 2016. A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow. arXiv:1602.09000 [physics].

Hartigan, J.A., 1975. Clustering Algorithms. Wiley.

IHSI, Mar. 2015. Population totale de 18 ans et plus menages et densites estimes en 2015. Institut Haitien de Statistique et D'Informatique.

Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transport. Res. C Emerg. Technol. 40, 63–74.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., Jun. 2011. Identifying important places in people's lives from cellular network data. In: Pervasive Computing. Springer, Berlin, Heidelberg, pp. 133–151.

Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F., Nov. 2012. Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PLoS One 7 (11), e49171.

JICA, 2013. Nairobi Personal Travel Survey. Japan International Cooperation Agency.

Kujala, R., Aledavood, T., Saramäki, J., Mar. 2016. Estimation and monitoring of city-to-city travel times using call detail records. EPJ Data Sci. 5 (1), 6.

Lozano-Gracia, N., Garcia Lozano, M., 2017. Haitian Cities: Actions for Today with an Eye on Tomorrow. Tech. rep.. World Bank, Washington, DC, USA.

Lu, X., Wetter, E., Bharti, N., Tatem, A.J., Bengtsson, L., Oct. 2013. Approaching the limit of predictability in human mobility. Sci. Rep. 3.

Nöel, R., Nov. 2012. Reconstruction et environnement dans la région métropolitaine de Port-au-Prince: cas de canaan ou la naissance d'un quartier ex-nihilo.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.-L., Sep. 2015. Returners and explorers dichotomy in human mobility. Nat. Commun. 6, 8166.

Prud'homme, R., Kopp, P., Nov. 2011. Urban transport in Port-au-Prince. Tech. rep.. IADB.

PTUMA, 2010. Encuesta de movilidad domiciliaria 2009-2010: movilidad en el área metropolitana de Buenos Aires. Proyecto de Transporte Urbano en Áreas Metropolitanas Argentinas.

Pulselli, R.M., Romano, P., Ratti, C., Tiezzi, E., 2008. Computing urban mobile landscapes through monitoring population density based on cell-phone chatting. Int. J. Des. Nat. Ecodyn. 3 (2), 121–134.

Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D., Sep. 2006. Mobile landscapes: using location data from cell phones for urban analysis. ResearchGate 33 (5), 727–748.

Reades, J., Calabrese, F., Ratti, C., Oct. 2009. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. Environ. Plan. B Plan. Des. 36 (5), 824–836.

Schneider, C.M., Belik, V., Couronne, T., Smoreda, Z., Gonzalez, M.C., May 2013. Unravelling daily human mobility motifs. J. R. Soc. Interface 10 (84), 20130246.

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., Feb. 2010. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.

Tatem, A.J., Jan. 2017. WorldPop, open data for spatial demography. Sci. Data 4, 170004.

Tatem, A.J., Qiu, Y., Smith, D.L., Sabot, O., Ali, A.S., Moonen, B., Dec. 2009. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malar. J. 8 (1), 287.