Comparison of User Presence Information from Mobile Phone and Sensor Data

Solohaja Rabenjamina, Razvan Stanica, Oana Iova, Hervé Rivano

Comparison of User Presence Information from Mobile Phone and Sensor Data

SOLOHAJA RABENJAMINA and RAZVAN STANICA, Univ Lyon, Inria, INSA Lyon, CITI, France
OANA IOVA and HERVÉ RIVANO, Univ Lyon, Inria, INSA Lyon, CITI, France

Data collected from mobile phones or from motion detection sensors are regularly used as a proxy for user presence in networking studies. However, little attention was paid to the actual accuracy of these data sources, which present certain biases, in capturing actual human presence in a given geographical area. In this work, we conduct the first comparison between mobile phone data collected by an operator and human presence data collected by motion detection sensors in the same geographical area. Through a detailed spatio-temporal analysis, we show that a significant correlation exists between the two datasets, which can be seen as a cross validation of the two data sources. However, we also detect some significant differences at certain times and places, raising questions regarding the data used in certain studies in the literature. For example, we notice that the most important daily mobility peaks detected in mobile phone data are not actually detected by on ground sensors, or that the end of the work-day activities in the considered area is not synchronised between the two data sources. Our results allow to distinguish the metrics and the scenarios where user presence information is confirmed by both mobile phone and sensor data.

CCS Concepts: • **Networks** → **Network mobility**; *Network measurement*.

Additional Key Words and Phrases: mobility models; user presence; mobile phone data; motion detection sensors

1 INTRODUCTION

Accurate mobility models, providing realistic user presence information at a given space and time, are essential in networking research [1, 2]. These models are generally based on some kind of real data, used to extract important mobility properties. Two main data sources have been used in the literature recently to retrieve user presence and mobility information: mobile phone data collected by cellular operators [3] and presence detection sensor data [4]. The first can cover an important number of users (i.e. millions) over large geographical areas (i.e. country wide), while the second is generally focused on the study of a specific environment (i.e. a neighborhood or a building). However, as discussed below, both these data sources present some inherent biases, which are yet to be studied and understood. This raises questions regarding the quality of the mobility models produced using these methodologies.

Mobile phone data collected by network operators for management and billing purposes covers a large part of the population of a country, providing much broader insights than classical surveys [5]. As a result, this type of data has been used in a variety of projects, from urban planning [6] to migration pattern analysis [7] and from the study of major events [8] to the one of epidemics propagation [9]. All these studies make the inherent assumption that mobile phone data represents a natural proxy for human presence and mobility in the studied area. However, there are two important factors with an impact on the accuracy of mobile phone data. First of all, the network operators can only collect mobile phone data when the user is active on the network, i.e. the mobile phone is exchanging data with the network. Of course, the probability of using a mobile phone is not uniformly distributed in space and time, as it can be easily seen in the operator data, which classically shows a much more important network presence during day time compared to night time [10]. Second, individual user data presents serious privacy issues [11], which means that spatio-temporal aggregated datasets are generally used [12]. However, this aggregation also implies some assumptions, once again related to the fact that the user sampling is not uniform, e.g. some user movements might be observed by the network with a certain delay, or the presence of some users might not be accounted for during certain time intervals.

On the other hand, presence detection sensors are widely available on the market and they can be usually installed. They are based on different technologies, such as infrared, ultrasonic, magnetic, image processing, or ultra-wide band [13]. Different biases are also encountered in this case. First of all, presence detection senors have a limited sensing range, which means some users might not be observed, despite being present in the area of study. Second, most presence detection sensors can not distinguish between two different persons, which leads to errors when such sensors are used to count the number of users present in an area, as some of them might be counted twice. Finally, if they are not well calibrated, presence detection sensors can also detect non-human presence of some wild or domestic animals.

In this work, we deploy human motion detection sensors in an area where we also have access to aggregated mobile phone data. We conduct a spatial analysis, where we deploy the sensors in nine different locations for a few weeks within the studied area, and a temporal analysis, where we collect 6 months of data in two specific locations. By comparing the sensor data with mobile phone data, we show a significant correlation between the two time series. Since the biases of the two data sources are different, we believe this correlation is a strong sign of the accuracy of the two datasets, which cross validate each other. However, we also observe that the correlation is less important at certain times and places. For example, the most significant daily peaks are very different in the two datasets, and the end of activities during the work-days seems particularly unsynchronised between the two data sources. In fact, important differences between sensor data and mobile data can be noticed during evening hours, as well as on Sundays. This indicates that user presence data produced from these data sources should be manipulated with precaution, as they can represent a relatively biased human mobility proxy in certain situations.

The remainder of this paper continues with Section 2 discussing related works. Section 3 details the two datasets analysed in this study, while Section 4 presents the metrics we used to compare them. Section 5 presents results related to a sensor deployment campaign covering nine different locations in the studied area for periods of a few weeks. These results are complemented by Section 6, where 6 months of data are collected in two specific locations and analysed. Finally, Section 7 concludes the paper.

2 RELATED WORK

The bias introduced by mobile phone data in the reconstruction of individual trajectories was characterised a decade ago by Ranjan *et al.* [14]. The authors showed that significant locations for the user (such as his home or work place) are correctly retrieved by using mobile phone data, but that the resulting trajectories present significant errors. This is a consequence of the spatio-temporal granularity of mobile phone data, where the user location is associated to the position of his serving base station and the location sampling is irregular and dictated by the phone activity. Caceres *et al.* [15] showed that mobile phone data misses short trips, but it still can be used to obtain accurate origin-destination matrices, while Qiu *et al.* [16] collected vehicular user data to prove that mobile phone data estimates the travel time of a car with an error of 5-15%. A series of works have been conducted in order to solve this individual trajectory problem, by using complex interpolation techniques [17] or historical user data [18].

However, individual user mobile phone data is rare and strongly regulated by privacy protection policies [11]. For these reasons, most of the mobile phone data used in the literature takes the form of aggregated datasets, indicating simply the number of users present in an area, or the number of users moving between two areas in a given time interval. Among other use cases, this type of data was heavily used recently to analyze human mobility during the Covid-19 pandemics [9]. Aggregated mobile phone data are regularly used as a proxy for human presence and mobility, for example to guide the evolution of the road infrastructure in a city [6] or to evaluate the potential benefits of a

ride-sharing service [19]. The conclusions of these studies are deeply related to the assumption that mobile phone data gives an accurate picture of human presence and mobility.

Multiple independent studies tried to assess the accuracy of aggregated mobile phone data using census data as ground truth, showing significant matching between the night-time user distribution in mobile phone data (considered as a proxy for home locations) and national and local census data in different areas: Los Angeles and New York [20], Israel [21], Barcelona and Madrid [22], or Milan [23]. Regarding mobility information, Schneider *et al.* [24] show that mobility motifs extracted from mobile phone data correspond to those obtained by using population surveys. At a country level, mobile phone data was used to estimate the destination distribution of people who left the affected area in the months following the 2010 Haiti earthquake, showing a good matching with a large survey conducted by the United Nations [7]. However, some survey-based studies also noticed significant differences with mobile phone data. For example, Tizzoni et al. [25] identify an overestimation in commuting flows inferred from mobile phone data with respect to national census data in France, Spain and Portugal. In the same lines, Wesolowski *et al.* [26] discover that users who travel a lot also use their mobile phones more, introducing an overestimation in the average level of mobility of the population.

Compared with these previous works, we do not use census data over large areas (cities or countries), but focus instead on a neighborhood-size area, where we collect human presence sensor data. The study that is closest to ours is the one conducted by Ma *et al.* [27], who collect highway mobility data using a licence plate recognition system and find a strong correlation with mobile phone data. We also notice a strong correlation between sensor and mobile phone data, but also investigate more original metrics, which uncover a different picture.

3 DATASETS

Our objective is to assess the accuracy of aggregated mobile phone data and that of human presence detection sensor data by cross correlation. For this, we focus on one geographical area, for which we collect both mobile phone data and human presence detection data. We target a suburban area, with little residential buildings and mostly hosting industrial and commercial buildings. The choice of this area was guided by social acceptability constraints: the deployment of sensors to detect human presence was conducted by the city administration and the city population was informed through multiple channels. We note that we did not use any individual data for this study, and that the collection of aggregated data was conducted in cooperation with the local administration.

The area of study represents one mobile network cell of a nation-wide mobile operator. For this cell, we have access to aggregated mobile phone data for the period July 2020 - March 2021. This data gives, with a 30 minutes granularity, the number of users entering and exiting the area covered by the cell. The aggregated data is produced by the mobile operator following a methodology that is quite common in the datasets used in the literature [5]. Practically, a user is associated with the most recent serving base station he used. For a given 30 minutes interval, if a user connects to the serving base station covering the area of interest, after being previously associated with a different base station, a user entry in the area is counted. Reversely, if a user previously associated with the area of interest is now observed in a different cell, this counts as an exit from the area. In our study, we do not distinguish between entries and exits. As explained below, this is a consequence of the fact that our sensors do not measure the direction of travel. Therefore, we sum the number of entries in and exits from the area in one single measure: the number of users moving in the area of interest. The granularity of 30 minutes is clearly a limiting factor for the possible resolution of the study. However, most studies using mobile phone data (*e.g.* [6, 10]) use a temporal granularity of 30 minutes or more, which seems to be sufficient to describe the user mobility.

For the second dataset, we also deployed motion detection sensors in the area of study. We used the sensors developed by Dahan *et al.* [28] for detecting urban mobility, which were tested and validated in real deployments. The motion detection component of these sensors uses a technology based on passive infrared receiver (PIR) and can detect movements over a distance of up to 7 meters, with a detection angle of 110 degrees. Its output is a binary response when a thermal motion is detected, and it can detect both passing-by pedestrians, as well as vehicles. The sensors were deployed solely on lamp posts situated outdoor. Of course, we note an obvious bias through the fact that a detected vehicle might transport several persons. These sensors were deployed during two different periods: from July 2020 to September 2020 and from January 2021 to March 2021. In the first period, the sensors were deployed for relatively short time intervals (from a few days to a few weeks) in nine different locations in the area of interest, as shown in Fig. 1. During the second period, the sensors were deployed at only two locations, denoted as SJ161 and SJ214 in Fig. 1 (these two locations were also continuously monitored during the first time period).

Fig. 1. The nine locations where sensors were deployed.

We note that the two periods for which we collected sensor data present two significant differences. The first one is obvious, in terms of season: the first collection period covers the summer holidays (with a high impact on this mostly industrial and commercial area), while the second collection period covers winter months. The second major difference is related to the Covid-19 sanitary measures: no particular measure was active during the first period, while a lockdown at 19h was in place during the second collection period.

4 METRICS

In order to compare aggregated mobile phone data and human presence sensor data, we use a series of metrics, detailed in the following. We denote by $\mathcal{M}$ the mobile phone data time series, which can be divided in daily time series $\mathcal{M}_d$, which itself is formed of mobile data mobility measurements, m_i, recorded every 30 minutes. Similarly, for the sensor data, the time series at a given location l, denoted as $\mathcal{S}^l$, is divided in daily time series $\mathcal{S}^l_d$, formed of sensor data measurements aggregated over a 30 minutes interval, s^l_i.

4.1 Correlation

The most obvious metric to compare two time series is their correlation, already used in the vehicular-oriented study by Ma *et al.* [27], as discussed in Sec. 2. In this work, we use the Pearson correlation coefficient computed on a daily basis, which indicates the linear correlation between two daily time series. This coefficient, representing the ratio between the covariance of the two time series and the product of their standard deviations, is defined as follows:

$$\rho_{(\mathcal{M}_d, \mathcal{S}_d^l)} = \frac{cov(\mathcal{M}_d, \mathcal{S}_d^l)}{\sigma_{\mathcal{M}_d} \cdot \sigma_{\mathcal{S}_d^l}}, \tag{1}$$

A correlation coefficient close to 1 or -1 implies a strong relationship (positive or negative) between the two time series, whereas a $\rho_{(\mathcal{M}_d, \mathcal{S}_d^l)}$ close to 0 indicates that the two time series are unrelated and very different.

However, simply computing $\rho_{(\mathcal{M}_d, \mathcal{S}_d^l)}$ can not account for the fact that the two data sources might be slightly shifted in time. Indeed, a person entering or exiting the area might not be detected exactly at the same time by the mobile network and the motion detection sensors. For this reason, we also measure the Pearson correlation coefficient between the daily time series with a certain delay shift in one of the two time series. Practically, since our datasets have a 30 minutes granularity, when we introduce a delay of 30 minutes, we align m_i and s_{i-1}, when the delay is 1 hour, we align m_i and s_{i-2}, etc.

We denote the sensor collected time series at location l, shifted with a delay τ, as $\mathcal{S}_d^l(\tau)$. In this case, the Pearson correlation coefficient with a delay τ is defined as:

$$\rho^{\tau}_{(\mathcal{M}_d, \mathcal{S}_d^l(\tau))} = \frac{cov(\mathcal{M}_d, \mathcal{S}_d^l(\tau))}{\sigma_{\mathcal{M}_d} \cdot \sigma_{\mathcal{S}_d^l(\tau)}}. \tag{2}$$

Please note that the Pearson correlation coefficient is agnostic of the actual amplitudes of the two time series, hence reducing the bias introduced in the sensor data by vehicles possibly carrying multiple mobile users.

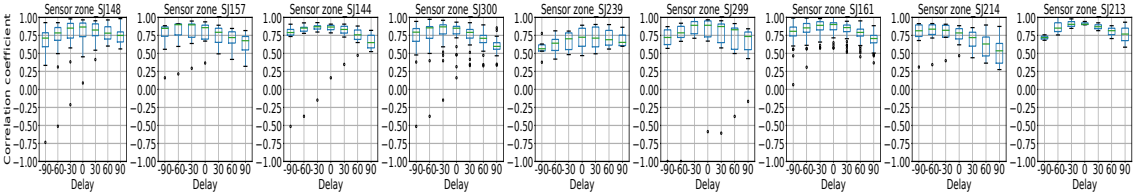

Fig. 2. Distribution of the Pearson correlation coefficient depending on the delay shift between the two time series.

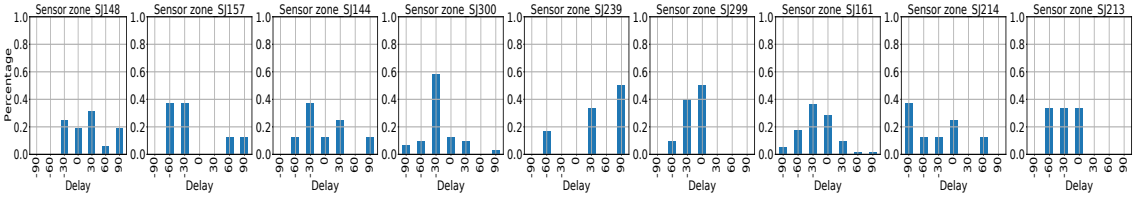

Fig. 3. Distribution of the daily delay shift between the time series giving the highest Pearson correlation coefficient.

4.2 Peak synchronisation

The detection of peaks in mobile phone data, and in mobility data in general, is a routine task. Peaks are often related to human activities, for example morning and evening mobility peaks are classically associated with commuting [25]. Similarly, detecting the location and the time of the most important mobility peaks in a given area is a common way to detect hotspots [29].

For this reason, we focus on peak synchronisation as a second metric in our study. First of all, we define a peak as a local maximum in a given time series, over a window of predefined size w. In other words, using the mobile phone data time series as an example:

$$P_{\mathcal{M}} = \{i | m_i \geq m_{i-k}, \forall k \in [-w, w]\} \tag{3}$$

In a similar way, we can define $P_{\mathcal{S}^l}$ as the peaks found in the sensor data at location l. We can now compute different peak synchronisation metrics, as follows. For each peak $p_m \in P_{\mathcal{M}}$, we compute $\delta_{p_m}(P_{\mathcal{S}^l})$ representing the time difference between p_m and the closest peak in the sensor data time series $\mathcal{S}^l$. Practically, if the peaks in the two time series are perfectly synchronised, $\delta_{p_m}(P_{\mathcal{S}^l}) = 0$. We note that the computation of this time difference between the peaks of two time series is not commutative, so we also compute $\delta_{p_s^l}(P_{\mathcal{M}})$. Finally, we compute $\delta_{p_m}(P_{\mathcal{S}})$, representing the time difference between p_m and the closest peak in all the sensor data time series, covering all the nine sensor locations. We also extract the two highest daily peaks in the different time series, denoted as $H_{\mathcal{M}}$ and $H_{\mathcal{S}^l}$, and compute the time difference between them using an analogous procedure to the one described above.

4.3 Start and end of the day

The night-time and day-time behavior is very different in mobile phone data [10]. Detecting home locations and work locations as the predominant user locations during night-time and, respectively, day-time, is standard procedure when working with mobile phone data [20, 21]. We therefore focus on detecting the start of the day (i.e. the beginning of human activities in the studied area) and the end of the day (i.e. the moment when most of the human activity in the area stops) in the two datasets.

To this end, we use the symmetric derivative, which is an approximation of the derivative on time series. Practically, we assume that the derivative of the time series is a good indicator for the start and the end of the day. More precisely, the start of the day is defined as the moment when the derivative has the highest value (we limit the studied interval for this purpose from 3 am to 12 pm), while the end of the day is defined as the moment when the derivative has the lowest value (we limit the studied interval for this purpose from 4 pm to 3 am).

This allows us to compute daily values for the start of the day in mobile phone data (denoted as $B_{\mathcal{M}}$) and in sensor data (denoted as $B_{\mathcal{S}^l}$). Similarly, we identify the end of the day $E_{\mathcal{M}}$ and $E_{\mathcal{S}^l}$. With this, we can compute δ_B and δ_E, the time difference between the start of the day, and respectively the end of the day, in the two data sources.

5 SPATIAL ANALYSIS

In this section, we discuss results related to the 3 months deployment in the summer 2020, using nine different sensor locations in the area of study, as shown in Fig. 1. To this end, we compare mobile phone data and sensor collected data by using the metrics described in Section 4.

5.1 Correlation coefficient

Fig. 2 shows, for each of the nine sensor locations, the distribution of their daily correlation coefficient with mobile phone data, depending on the delay shift between the two time series. We test several values for this delay, from -90 minutes to +90 minutes, with a step of 30 minutes.

We can see from the boxplots that, in general, at every location there is at least one delay shift where the median correlation coefficient is superior to 0.8, a sign of significant correlation between the two data sources. The only exception comes from location SJ239, which shows a much lower Pearson correlation coefficient compared to the other locations. As it can be seen in Fig. 1, in this case the sensor is deployed in the parking place of a commercial area. Our intuition in this case is that the parking place has slightly different dynamics than the rest of the area of study, as people can enter and exit the area without necessarily visiting the parking place.

We also notice that the highest median correlation coefficient is usually obtained for a delay of -30 or 0 minutes, with specific patterns in SJ239 and SJ214. These fluctuations indicate that, depending on the location, a certain shift exists between the two data series. To better study this phenomenon, Fig. 3 shows the histogram of the delay shift producing the highest daily correlation in each of the nine sensor locations.

The first observation in Fig. 3 is that the delay giving the best correlation shows important daily variations. For example, in location SJ148, the delay between the two time series varies from -30 to +90 minutes. Only one location, SJ300, has a delay value (-30 minutes) achieving the highest correlation for more than 50% of the days. The only location where the mobile phone data and the sensor data seem strongly synchronised (i.e. a 0 delay) is SJ299, situated in the south of the area of study. Finally, the specific pattern of the parking place location SJ239 is again visible, with a +90 minutes delay usually giving the best results, but also with days where a totally opposite behavior (-60 minutes delay) is observed.

Overall, our results confirm that the aggregated mobile phone data is strongly correlated with the human mobility measured by the sensors. Yet, a delay often occurs between the two time series. This can be explained by the pre-processing done on the mobile operator side in order to produce the aggregated dataset. Indeed, the mobility of a user can be observed with a certain delay on the mobile network, since the user is not necessarily active on the mobile network when moving. Finally, some specific properties can be observed at different sensor locations, demonstrating the existence of a spatial diversity. However, it is striking to note the very high correlation between the two time series at practically all the locations, despite this spatial heterogeneity. We consider that this significant correlation allows for a cross correlation of the two data sources, showing that they accurately represent the user presence in the studied area.

5.2 Peaks synchronisation

We proceed by comparing the peaks in the mobile phone data and in the sensor data, using the metrics described in Section 4. Fig. 4 shows the distribution of $\delta_{p_m}(P_{\mathcal{S}})$. Practically, for every peak detected in the mobile phone data, we find the closest peak in the sensor data, in any of the nine locations. We can notice once again a strong correlation, with a probability of more than 95% to find a peak in at least one sensor location within less than 1h with respect to mobile phone data. The delay observed in terms of correlation is also visible here, with most of the peaks being shifted by 30 minutes between the two data sources.

A similar trend can be observed when we analyse the different locations individually, as shown by the distribution of $\delta_{p_m}(P_{\mathcal{S}^l})$ in Fig. 5. In this case, we also run a clustering algorithm on the results of the nine locations in order to automatically detect locations with similar trends. From a practical point of view, the results at each location are seen

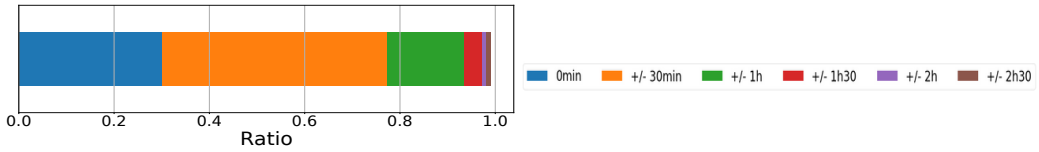

Fig. 4. Distribution of $\delta_{pm}(P_{\mathcal{S}})$.

as a vector of six elements and a hierarchical clustering algorithm is used to group the similar vectors together. The obtained classes are labeled as A-E in Fig. 5.

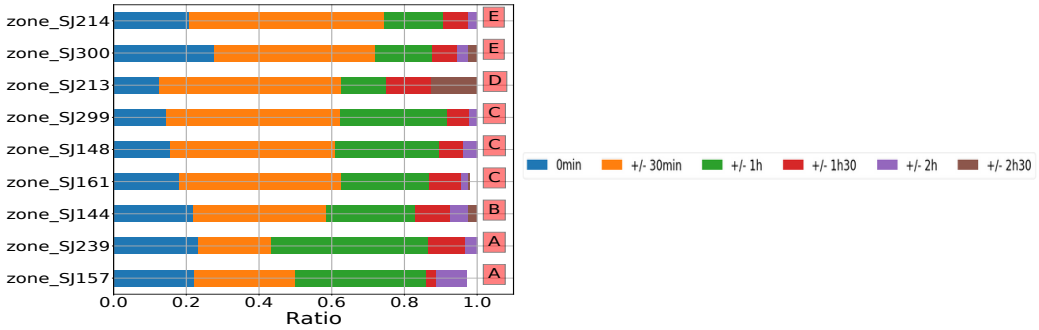

Fig. 5. Distribution and clustering of $\delta_{pm}(P_{\mathcal{S}^l})$.

For most of the locations, the probability to find a peak in less than 1h with respect to the mobile phone data is higher than 0.8. The only exception is SJ213, which shows a more reduced peak synchronisation level. For locations clustered in classes B to E, the most likely delay between the peaks is of 30 minutes. However, a different pattern can be noticed for locations in class A, where a delay of 1h between the peaks is more probable. With respect to the clustering, the similarity in terms of peak synchronisation behavior does not seem to depend on geographical proximity. In fact, except for SJ148 and SJ161, clustered together in class C, the locations classified together are not relatively close geographically.

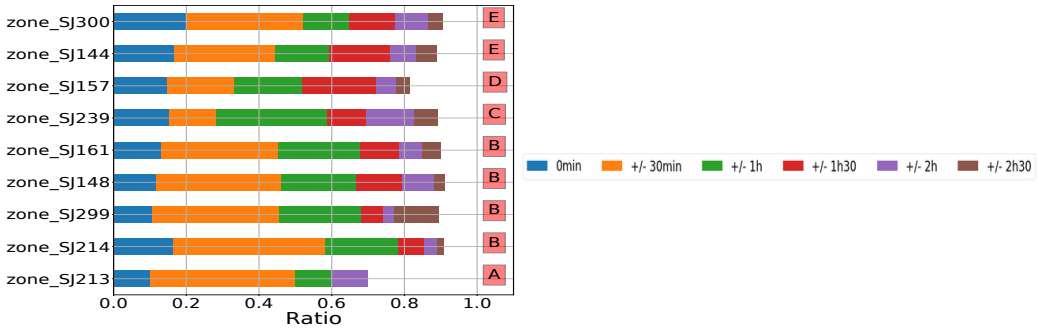

Fig. 6. Distribution and clustering of $\delta_{p_s^l}(P_{\mathcal{M}})$.

The distribution of $\delta_{p_s^l}(P_{\mathcal{M}})$, shown in Fig. 6, gives a different view. We recall that, in this case, we first detect the peaks in the sensor data and then find the closest match in the mobile phone data. The synchronisation between the

peaks is much weaker in this sense, with a probability of around 0.6 of finding a corresponding peak in less than 1h. For the location SJ213, the delay between the peaks is higher than 2h30 (the limit we set in Fig. 6) in more than 30% of the cases. These results are a consequence of the sensor data being more dynamic and presenting more peaks than the mobile phone data. Practically, whenever a peak shows up in the aggregated mobile phone data, it is likely that it also appears in the sensor data. However, local peaks in the sensor data are less likely to appear in the more spatially aggregated mobile phone data.

Since the number of peaks is different in the two data sources, we focus our analysis on the two most important daily peaks in each dataset. We choose this value because, generally, there are two main daily peaks observed in mobile phone data [10]. As detailed in Section 4, for each day we extract $H_{\mathcal{M}}$, the two largest peaks in the mobile phone data. We also extract $H_{\mathcal{S}}$, the two largest peaks in the sensor data, considering all the locations. The time difference between the peaks in $H_{\mathcal{M}}$ and those in $H_{\mathcal{S}}$ is presented in Fig. 7, where we distinguish the results for the first peak and the second peak in $H_{\mathcal{M}}$.

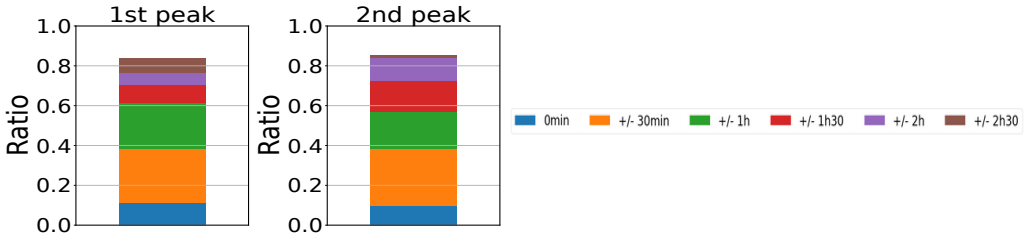

Fig. 7. Time difference between the peaks in $H_{\mathcal{M}}$ and those in $H_{\mathcal{S}}$.

We notice that the time difference between the highest daily peaks in the two datasets is lower than 1h only 60% of the time. Moreover, almost 20% of the time, this time difference is higher than 2h30. The results are even more surprising when looking at the nine locations individually, in Fig. 8. With the exception of SJ213, where the peaks are very often in a 30 minutes range, all the other locations show a rather poor synchronisation with mobile phone data in terms of the most important peaks. The probability to have a time difference lower than 1h for the most important peak is generally below 0.5, while it is a little bit higher for the second most important daily peak. We also run the clustering algorithm on the nine locations based on these results, obtaining four classes (A-D) for the most important daily peak and three classes (1-3) for the second most important one.

These results indicate significant differences between the most important peaks in the mobile phone data and those detected by human presence sensors in the area of study. These peaks are rarely synchronised, as it can be seen by the reduced amount of blue color in Fig. 7 and in Fig. 8. Moreover, even when considering a 1h delay between the peaks in the two datasets, it is more likely not to find the peaks detected by the sensors in the mobile phone data. This raises doubts on the results of studies detecting significant peaks in mobile phone data, e.g. those focused on hotspots [29], since these peaks are not confirmed by our second data source.

5.3 Start and end of the day

We wrap up this spatial analysis by comparing the start of the day and end of the day in the two datasets, following the methodology described in Section 4. Fig. 9 shows the synchronisation of the start of the day between the mobile phone data and the sensor data. We can see a good synchronisation, as all the nine locations have at least 80% of their days

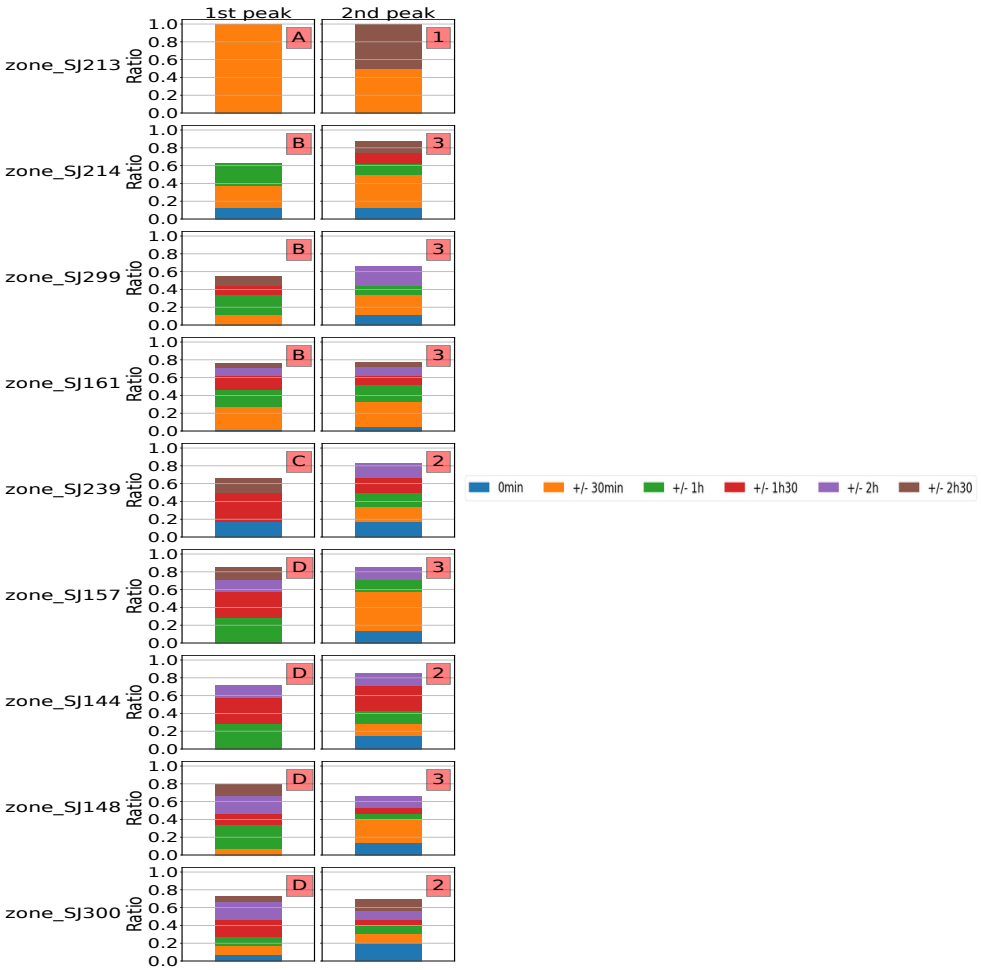

Fig. 8. Time difference between the peaks in H_{S^l} and those in $H_{\mathcal{M}}$.

with an 1h synchronisation with the mobile data. Some locations show even a better synchronisation: SJ148, SJ299 and SJ213 have a 30 minutes synchronisation probability above 80%.

On the other side, the synchronisation of the end of the day, presented in Fig. 10, is much more problematic. The end of day synchronisation within an 1h interval has a probability of around 60% for most locations, with some locations, such as SJ299 or SJ157, reaching even lower results. Even considering the 2h30 limit used in our tests, the difference is more significant for around 15% of the days. As we will discuss later, this difference usually comes from a more significant evening activity detected by the sensors, which does not appear in the mobile phone data

These results confirm an excellent correlation between the two data sources in the morning hours, after accounting for the 30 minutes delay between the datasets that was observed by all our metrics. However, significant differences are noticed between mobile phone data and sensor data with respect to the end of the day, where the synchronisation

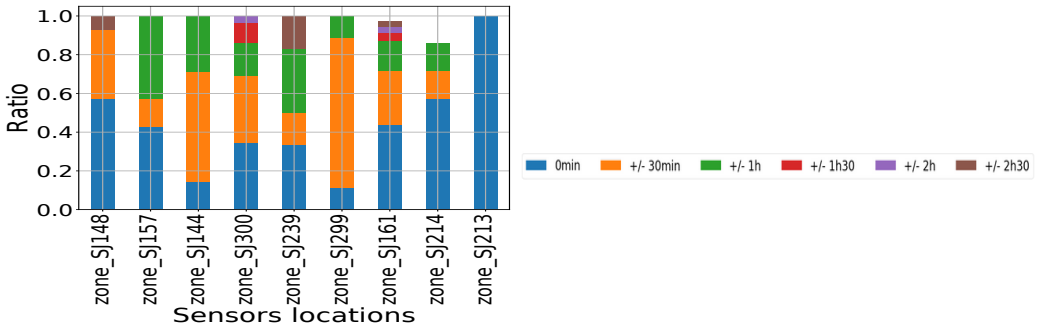

Fig. 9. Start of the day time difference between mobile data and sensors

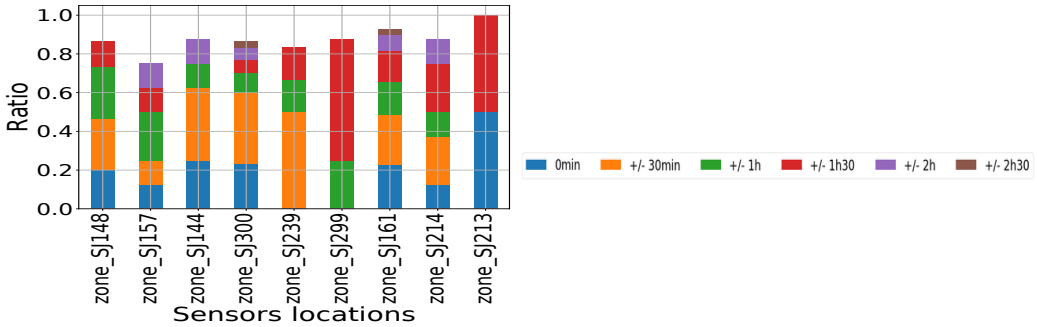

Fig. 10. End of the day time difference between mobile data and sensors

between the two data series is weak. Studies such as those focusing on the evening commuting behavior [25] should be aware of these differences.

6 TEMPORAL ANALYSIS

In this section, we complement the summer 2020 results discussed in Section 5 with 3 months of data collected during winter 2021. For this period, we only collect data for two sensor locations: SJ161 and SJ214. With 6 months of data for these two locations, we conduct an in-depth temporal analysis.

6.1 Correlation coefficient

We represent the Pearson correlation coefficient per day of the week, for the two locations and the two time periods, in Fig. 11. The figure only presents results for Monday, Thursday and Sunday, all the other days (including Saturday) being very similar to the Monday results.

The results confirm the very strong correlation between the two time series, with a correlation coefficient superior to 0.9. There are, however, two exceptions to this. Thursdays during the summer period at location SJ214 shows a median correlation coefficient below 0.75, a phenomenon no longer observed during the winter period. The second exception is represented by Sundays, when the correlation is much lower compared to the other days, especially during the summer period. We also observe the delay shift of 30 minutes between the two series, already noticed in Section 5.

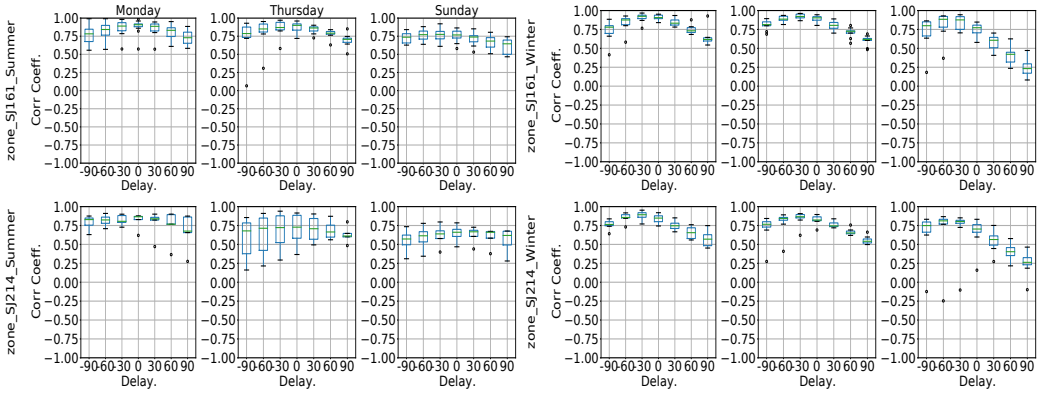

Fig. 11. Distribution of the Pearson correlation coefficient per day of the week.

6.2 Peaks synchronisation

Regarding the peaks, Fig. 12 and Fig. 13 show the distribution of $\delta_{pm}(P_{\mathcal{S}^l})$ per day of the week, for the two studied periods. We only show results for location SJ161, since the results for SJ214 are very similar. The main observation here is that the synchronisation of the peaks degrades during the winter period. For most of the days of the week (the only exception is the Tuesdays), the peak synchronisation probability with a 1h window reduces by 10% or more. We can also notice that, during both periods, the peak synchronisation is weaker towards the end of the week, and especially on Sundays during winter.

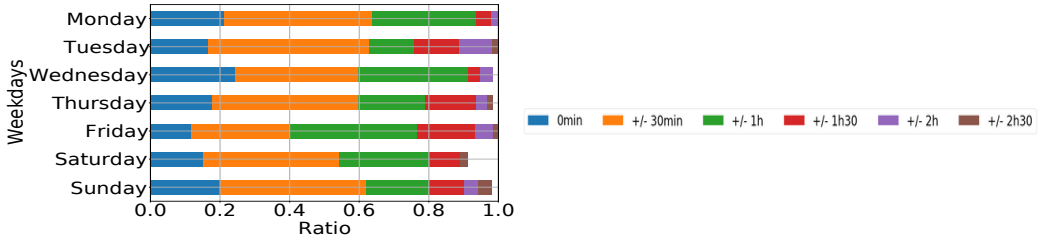

Fig. 12. Distribution of $\delta_{pm}(P_{\mathcal{S}^l})$ for location SJ161 during summer.

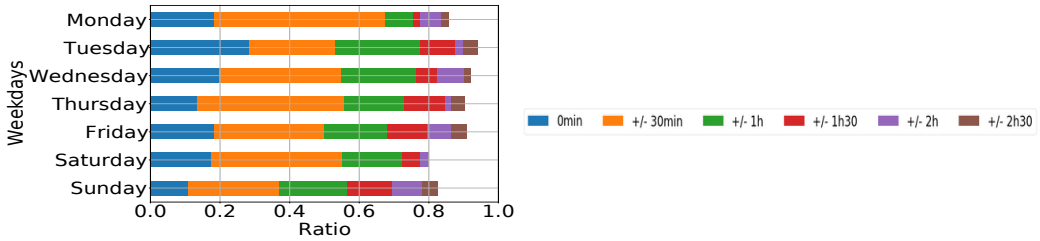

Fig. 13. Distribution of $\delta_{pm}(P_{\mathcal{S}^l})$ for location SJ161 during winter.

Focusing on the most important peaks, Fig. 14 shows the distribution of the most significant daily peak as a function of the hour of the day. We observe that data from the two locations is very consistent, and the peaks in the sensor data are much more uniformly spread over the entire day, while the mobile phone data generally presents peaks in the morning (8h30 - 9h) and the evening (17h30 - 18h). This underlines, once again, the poor correlation of the two data sources in detecting mobility peaks.

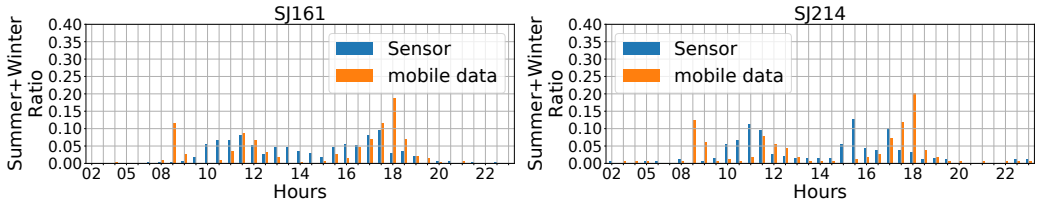

Fig. 14. Hourly distribution of the most important peaks in mobile phone data and in sensor data.

However, we notice an important exception, presented in Fig. 15. The distribution of the most significant peaks on Saturdays during the winter period is very similar in the two datasets, as shown in the figure for location SJ161.

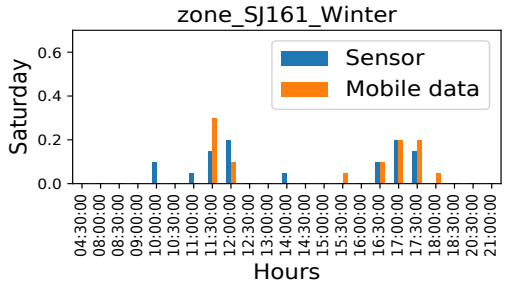

Fig. 15. Most important daily peak at location SJ161 during the winter period.

6.3 Start and end of the day

As discussed in Sec. 5, the start of day generally presents a good synchronisation in the two datasets. However, as shown in Fig. 16, some days of the week have a lower synchronisation than others, e.g. Saturdays for SJ161 and Mondays for SJ214. It is difficult to distinguish a general trend here, but overall the results seem to improve during the winter period.

This improvement is obvious for the end of the day results, in Fig. 17, where the synchronisation is much more important during the winter, compared with the summer period. The explanation for this behavior is that, during the winter period, a local lockdown at 19h was in place because of the Covid-19 pandemic. Therefore, the evening mobility observed in the sensor data during summer, which had a significant impact on the end of the day results, was no longer present during the winter time.

7 CONCLUSION

In this work, we assess the accuracy of user presence data used in recent mobility modelling solutions, by comparing aggregated mobile phone data with data collected by motion detection sensors. We define original metrics to compare

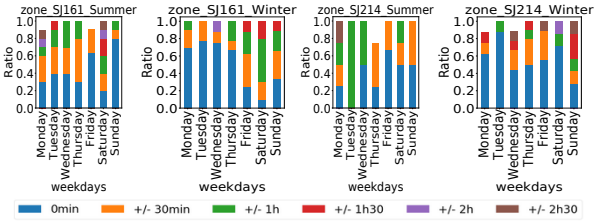

Fig. 16. Synchronisation of the start of the day depending on the weekday between sensor and mobile data .

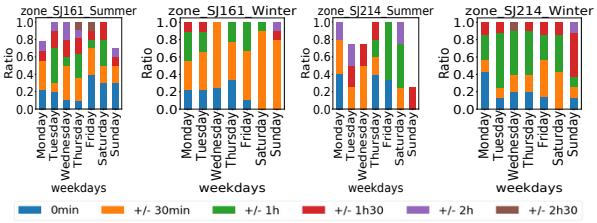

Fig. 17. Synchronisation of the end of the day depending on the weekday between sensor and mobile data .

the two data sources and our results indicate that there is indeed a strong correlation between the two datasets, which validates the realism of the data. However, we also notice some important differences, which need to be accounted for in future studies based on mobile phone data: *i)* there is generally a delay in the range of 30 min between the two time series; *ii)* the most important peaks in the two time series are weakly synchronised, with mobile phone data apparently biased by the user activity on the mobile network; *iii)* the correlation between the two data sources is less important on evenings and on Sundays. Nevertheless, a future study on other areas is required in order to generalize the behaviors we have observed in this study.

REFERENCES

[1] A. Jardosh, E. Belding-Royer, K. Almeroth, S. Suri, "Towards Realistic Mobility Models for Mobile Ad Hoc Networks", *Proc. ACM MobiCom*, San Diego, CA, USA, Sep. 2003.

[2] M. Musolesi, C. Mascolo, "Designing Mobility Models based on Social Network Theory", *ACM Mobile Computing and Communications Review*, vol. 11, no. 3, Jul. 2007.

[3] K. Xu, R. Singh, M. Fiore, M. Marina, H. Bilen, M. Usama, H. Benn, C. Ziemlicki, "SpectraGAN: Spectrum based Generation of City Scale Spatiotemporal Mobile Network Traffic Data", *Proc. ACM CoNEXT*, virtual event, Dec. 2021.

[4] Y. Wang, A. Yalcin, C. Vandeweerd, "An Entropy-based Approach to the Study of Human Mobility and Behavior in Private Homes", *Plos One*, vol. 15, no. 12, Dec. 2020.

[5] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Large-scale Mobile Traffic Analysis: A Survey", *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, Jan. 2016.

[6] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, L. Sbodio, "AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data", *Proc. ECML PKDD*, Prague, Czechia, Sep. 2013.

[7] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, J. von Schreeb, "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti", *PLoS Medicine*, vol. 8, no. 8, Aug. 2011.

[8] B. Cici, M. Gjoka, A. Markopoulou, C.T. Butts, "On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology", *Proc. ACM MobiHoc*, Hangzhou, PRC, Jun. 2015.

[9] K. Grantz, H. Meredith, D. Cummings, J. Metcalf, B. Grenfell, J. Giles, S. Mehta, S. Solomon, A. Labrique, N. Kishore, C. Buckee, A. Wesolowski, "The Use of Mobile Phone Data to Inform Analysis of COVID-19 Pandemic Epidemiology", *Nature Communications*, vol. 11, no. 4961, Sep. 2020.

[10] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, Z. Smoreda, "A Tale of Ten Cities: Characterizing Signatures of Mobile Traffic in Urban Areas", *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, Oct. 2017.

[11] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. Le Hello, U. Aivodji, B. Olivier, T. Quertier, R. Stanica, "Privacy in Trajectory Micro-Data Publishing: A Survey", *Transactions on Data Privacy*, vol. 13, Apr. 2020.

[12] S. Uppoor, C. Ziemlicki, S. Secci, Z. Smoreda, "On Mobile Traffic Distribution over Cellular Backhauling Network Nodes", *Proc. IEEE CCNC*, Las Vegas, USA, Jan. 2016.

[13] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, "Towards a Sensor for Detecting Human Presence and Characterizing Activity", *Energy and Buildings*, vol. 43, no. 2, Feb. 2011.

[14] G. Ranjan, H. Zang, Z.-L. Zhang, J. Bolot, "Are Call Detail Records Biased for Sampling Human Mobility?", *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, Jul. 2012.

[15] N. Caceres, L. Romero, F. Benitez, "Exploring Strengths and Weaknesses of Mobility Inference from Mobile Phone Data vs. Travel Surveys", *Transportmetrica A: Transport Science*, vol. 16, Feb. 2020.

[16] Z. Qiu, J. Jin, P. Cheng, B. Ran, "State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information Systems", *Proc. TRB 86th Annual Meeting*, Washington, USA, Jan. 2007.

[17] G. Chen, A. Carneiro Viana, M. Fiore, C. Sarraute, "Complete Trajectory Reconstruction from Sparse Mobile Phone Data", *EPJ Data Science*, vol. 8, no. 30, Oct. 2019.

[18] L. Bonnetain, A. Furno, N.-E. El Faouzi, M. Fiore, R. Stanica, Z. Smoreda, C. Ziemlicki, "TRANSIT: Fine-grained Human Mobility Trajectory Inference at Scale with Mobile Network Signaling Data", *Transportation Research Part C: Emerging Technologies*, vol. 130, no. 103257, Sep. 2021.

[19] B. Cici, A. Markopoulou, E. Frias-Martinez, N. Laoutaris, "Quantifying the Potential of Ride-Sharing using Call Description Records", *Proc. ACM HotMobile*, Jekyll Island, USA, Feb. 2013.

[20] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, "Ranges of Human Mobility in Los Angeles and New York", *Proc. IEEE PerCom*, Seattle, USA, Mar. 2011.

[21] S. Bekhor, Y. Cohen, C. Solomon, "Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions", *Journal of Advanced Transportation*, vol. 47, no. 4, Jun. 2013.

[22] M. Lenormand, M. Picornell, O. Cantu-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, J. Ramasco "Cross-Checking Different Sources of Mobility Information", *PLoS ONE*, vol. 9, no. 8, Aug. 2018.

[23] R. Douglass, D. Meyer, M. Ram, D. Rideout, D. Song, "High Resolution Population Estimates from Telecommunications Data", *EPJ Data Science*, vol. 4, May 2015.

[24] C.M. Schneider, V. Belik, T. Couronne, Z. Smoreda, M.C. Gonzalez, "Unravelling Daily Human Mobility Motifs", *Journal of the Royal Society Interface*, vol. 10, no. 84, May 2013.

[25] M. Tizzoni, P. Bajardi, A. Decuyper, G.K.K. King, C.M. Schneider, V. Blondel, Z. Smoreda, M.C. Gonzalez, V. Colizza, "On the Use of Human Mobility Proxy for the Modeling of Epidemics", *PLoS Computational Biology*, vol. 10, no. 7, Jul. 2014.

[26] A. Wesolowski, N. Eagle, A. Noor, R. Snow, C. Buckee, "The Impact of Biases in Mobile Phone Ownership on Estimates of Human Mobility", *Journal of the Royal Society Interface*, vol. 10, no. 81, Feb. 2013.

[27] J. Ma, H. Li, F. Yuan, T. Bauer, "Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data", *International Journal of Transportation Science and Technology*, vol. 2, no. 3, Sep. 2013.

[28] M. Dahan, A.A. Mbacké, O. Iova, H. Rivano, " Challenges of Designing Smart Lighting", *Proc. EWSN*, Lyon, France, Feb. 2020.

[29] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, G. Pujolle, "Estimating Human Trajectories and Hotspots through Mobile Phone Data", *Computer Networks*, vol. 64, May 2014.