Performance of home detection from mobile phone data

Maarten Vanhoof*[1], Clement Lee†[2] and Zbigniew Smoreda‡[3]

[1]Centre for Advanced Spatial Analysis, The Bartlett, University College of London
[2]Open Lab, School of Computing Science, Newcastle University
[3]SENSe department, Orange Labs, France

January 31, 2019

Summary

Mobile phone data form a promising data source. In many analyses, a prerequisite step to their deployment is the detection of home locations from individual users. Yet, little research exists on validation or uncertainty estimation of home detection methods. We present an empirical analysis of home detection methods for a French dataset. We analyze the validity of 9 Home Detection Algorithms (HDAs), and assess different sources of uncertainty. Findings show that performance of home detection is moderate and space-time dependent, that the duration of observation has a clear effect, and that the influence of criteria and parameter choice is small.

KEYWORDS: Mobile phone data, home detection, validation, error estimation.

1 Introduction

There exists an underdeveloped quality assessment of home detection methods for geo-located traces that are captured on a non-continuous base (Vanhoof et al., 2018a; Bojic et al., 2015). This is rather surprising as the use of such traces in analysis is almost always preceded by a methodological step that seeks to define a *home* for individual users. Take for example the case of Call Detailed Record (CDR) data, a type of mobile phone data, where location of users is collected every time a user makes a call or text on the operator's network. Detecting home locations from CDR data for individual users is important for different strands of mobile phone data research. To infer long-distance travel, tourism trips, or commuting, for example, home detection is needed in order to determine when a user is performing a specific type of mobility or not (Vanhoof et al., 2017; Janzen et al., 2018). In order to link mobile phone data to other data sources, such as census data, home detection is a prerequisite step too. Indicates derived from mobile phone data, for example, need to be aggregated in space to become comparable with, for example, census data (Cottineau and

* m.vanhoof@ucl.ac.uk

† clement.lee@newcastle.ac.uk

‡ zbigniew.smoreda@orange.com

Vanhoof, 2019; Pappalardo et al., 2016; Vanhoof et al., 2018b). These aggregations, are typically done using home detection, meaning mobile phone users get aggregated based on their assumed home locations.

The goal of this paper is to empirically explore the nation-wide performance of home detection methods for a case study in France and with a focus on the sensitivity to user choices such as the HDA-choice, the chosen period of observation and the chosen duration of observation. A clearer insight into the combined effects of user choices on the quality of home detection is desirable, and can help future work when making user choice as well as it can inform on the uncertainty and error related to home detection methods when performed on CDR data.

2 Detecting home locations from CDR data

2.1 The French CDR dataset

In our analysis, we will use an anonymized mobile phone dataset recorded by the French Operator Orange. The dataset covers mobile phone usage of 18 million subscribers on the Orange network in France during a period of 154 consecutive days in 2007 (May 13, 2007 to October 14, 2007). Mobile phone penetration is being estimated at 86% at that time and given a population of 63.945 million inhabitants during the observed period, that is about 32% of all French mobile phone users and 28% of the total population.

The mobile phone dataset consists of Call Detailed Record (CDR) data, which are typically collected by mobile phone service providers for billing and network maintenance purposes. Every time a call or text is initiated or received, CDR data store locational (the used cell tower), temporal (time and duration of usage), and interactional (who contacts whom) information for both correspondents. Location traces from CDR data thus are non-continuous as they are user initiated and rather sparse in time. In compliance with ethical and privacy guidelines CDR data are anonymized.

2.2 Defining nine simple home detection algorithms

Most HDAs that are deployed on CDR data consist of single-step approaches, which detect a *home* by selecting the cell tower that accords best to an imposed decision rule. This is opposed to two-step approaches where spatial grouping of cell towers is performed as an extra step. The decision rules applied in HDAs can be simple or complex, meaning that they are based on one criterion or several criteria, respectively (Vanhoof et al., 2018a). We opt to use simple decision rules over complex decision rules, as this allows better singling out the effect of criteria choice. The used criteria are given in Table 1. Clearly, some of these criteria are subject to a parameter choice. For example, in the case of the nighttime criteria, a definition of nighttime has to be specified such as between 21.00 and 07.00 hours.

Inspired by Vanhoof et al. (2018a); Bojic et al. (2015), we define nine HDAs by combining different criteria and parameter choices. They are described in Table 1 and will be deployed in our analysis.

We will apply these HDAs to different time periods with different durations, in order to assess the influence on performance.

Table 1: Description of deployed HDAs

Criteria	**Parameters**	**Name**	**Description: 'home is cell tower where:'**
Maximum Amount	/	MA	Most activities occurred
Distinct days	/	DD	the maximum active days were observed
Time constraints	19,9	TC-19-9	Most activities occurred between 19pm and 9am (night-time)
Time constraints	19,9,weekend	TC-19-9-WE	... between 19pm and 9am (night-time) and during weekend days
Time constraints	21,7	TC-21-7	... between 21pm and 7am (night-time)
Time constraints	21,7,weekend	TC-21-7-WE	... between 19pm and 9am (night-time) and during weekend days
Time constraints	9,19	TC-9-19	... between 9am and 19pm (daytime)
Time constraints	9,19,week	TC-9-19-WK	... between 9am and 19pm (daytime) but only during weekdays
Time constraints	weekend	TC-WE	... during weekend days only (Sat and Sun)

2.3 Assessing home detection performance at nation-wide scale

Validation of HDAs at the individual level is not straightforward because collecting individual level ground truth data is extremely expensive and comes with increased privacy risks. As a consequence, researchers have to settle with high-level validation practices. In our analysis, we will use aggregated population counts from census data as a ground truth dataset to compare against aggregated user counts that are the results of deploying previously defined HDAs.

To measure the performance of the HDAs, we compare the outcome of each algorithm with the ground truth data. Specifically, we evaluate the degree of similarity between a vector of user counts (based on a HDA), denoted by $\vec{x}$, and a vector of population counts (based on census data), denoted by $\vec{y}$, both aggregated at cell tower level. Both vectors $\vec{x}$ and $\vec{y}$ thus have an equal length representing the 18,273 cell towers in the Orange network.

Because of the unknown spatial distribution of the 28% sample of Orange users, our assessment of similarity cannot be absolute. Therefore, we define performance measures based on the relative similarities between both vectors as can, for example, be done by calculating the Pearson correlation coefficient (Pearson's R) between vectors $\vec{x}$ (user counts) and $\vec{y}$ (population validation count):

$$Pearson's\ R(\vec{x},\vec{y}) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}} \tag{1}$$

Once a performance measure is calculated for all nine HDAs during all time periods, we can evaluate the influence of criteria choice, parameter choice, duration of observation and period of observation on performance.

3 Results and discussion

After running the nine HDAs on all users in the CDR dataset and for different periods, we find the Pearson correlation coefficients to range between 0.45 and 0.60, which indicates a moderate performance (figure 1). It forms an interesting observation that criteria and parameter choice are less influential compared to time period or, sometimes, even duration of observation choice. For periods with a 14-day duration, for example, the effect of criteria choice is about 0.025 (expressed in Pearson's R) whereas the summer period effect is about 0.15, or thus an order higher (figure 1 B).

Still, some interesting observations can be made when comparing the performance of different HDAs. One observation is that the TC criterion outperforms the MA and DD criteria for some parameters (such as the 19-9) but definitely not for all. In other words, parameter choice for the TC criterion does have an impact on performance. For example, defining nighttime between 21-7 hours instead of 19-9 hours results in substantial performance loss for all 14-day periods investigated (figure 1 C). Even more remarkable is that the 21-7 parameter, at least for 14 days durations, is consistently outperformed by the 9-19 parameter, which is a daytime definition (figure 1 C). This finding drastically challenges the assumption that using nighttime would be better because people are more at home then. The performance of different TC parameters is also influenced by the time period. Using nighttime and weekends, for example, outperforms using weekdays during non-summer periods, but the true for all 14-day duration periods in August (see figure 1 D).

The most remarkable finding is that it is strongly dependent on the duration of observation as is emphasised further in figure 2. More specifically, we find that HDAs using shorter durations of observations (such as the 14-day duration) perform amongst the worst when observations are made during summer but amongst the best when the observations are made outside summer months. This is in contrast to the month and 30-day durations, where performance is somewhere in between, depending on the proportion of the time period that is in July or August and independent from the deployed HDA. For the 154-day duration, we find that for some HDAs, the longer time period is capable of mitigating the effect of summer holidays, leading the performances similar to the best performance of shorter durations (for example the TC-19-9 and TC-WE algorithms). This however is not true for all HDAs. Performances for the 154-day duration of the MA and TC-9-19-WK algorithms, for example, are not even close to the performance levels of shorter duration periods that are not occurring in summer.

Figure 1: Performance over time of HDAs with different criteria and/or parameters for all time periods with a 14-day (A,B,C,D) and 154-day (E,F,G,H) duration of observation.

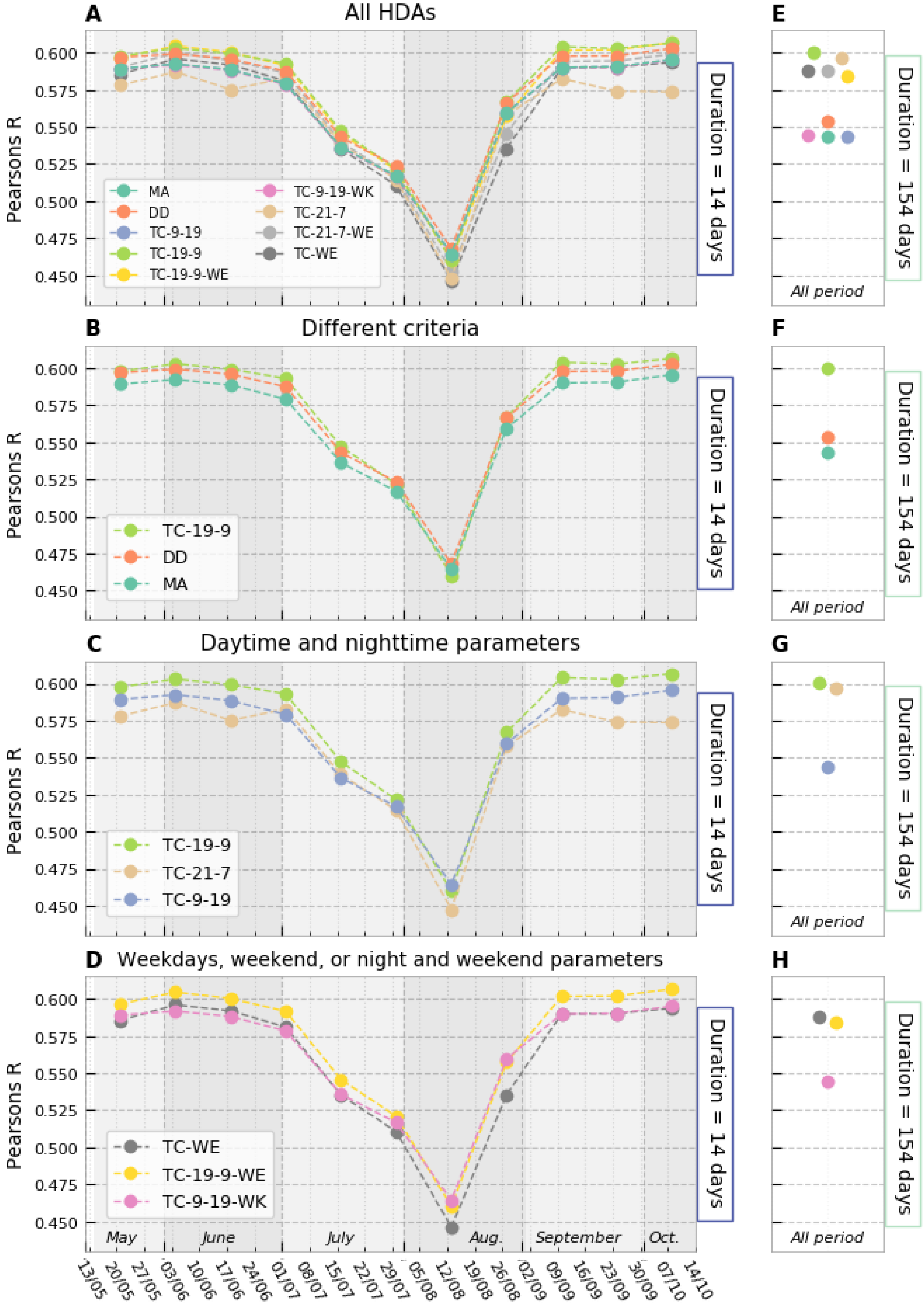
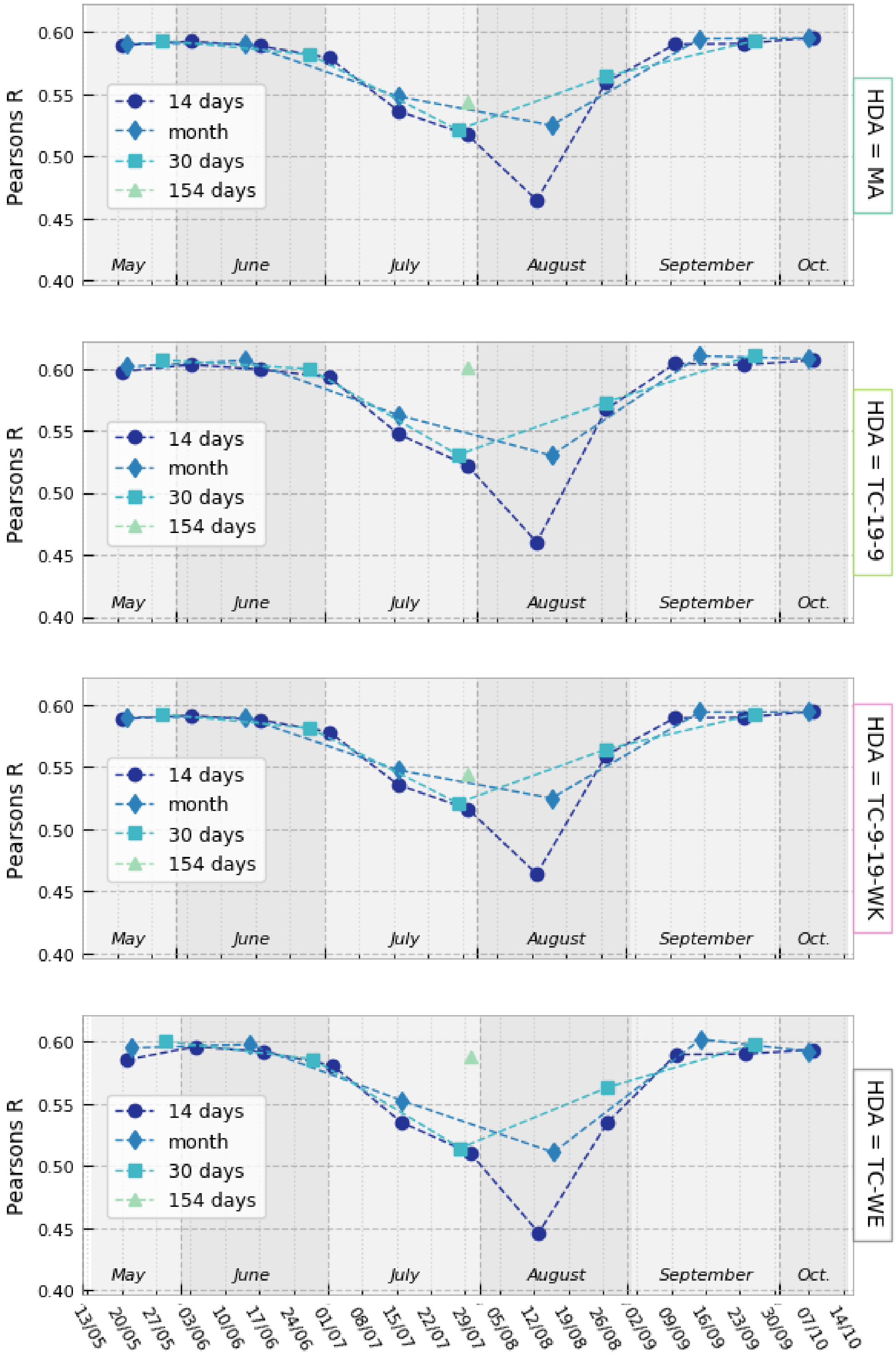


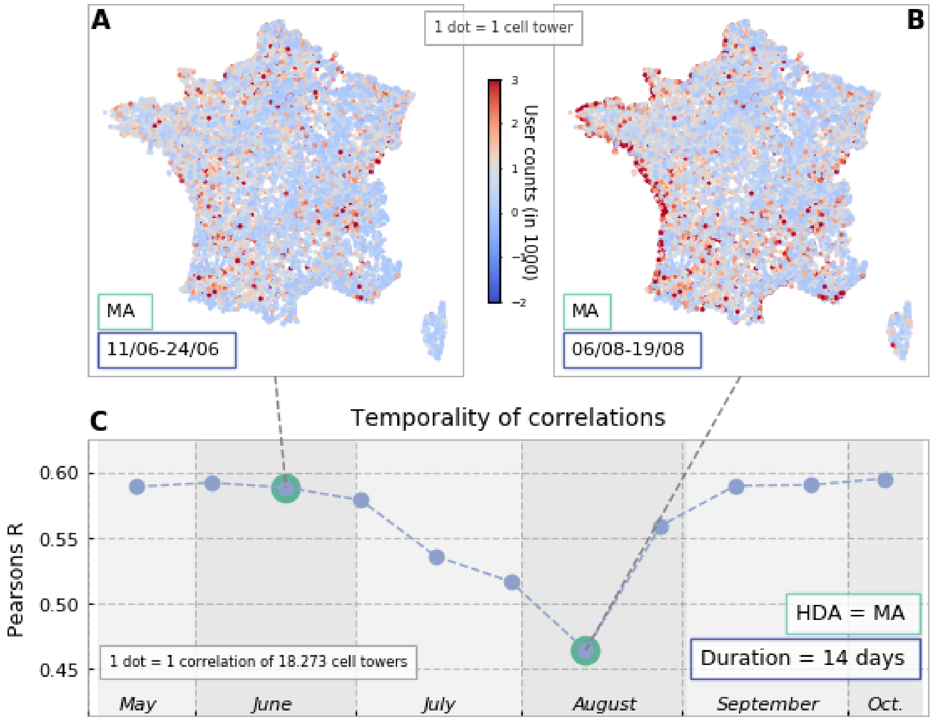
Figure 2: Performance over time of 4 algorithms for time periods characterized by different durations of observation.



In the case of home detection methods, there are several consequences to the absence of studies that assess validity, sensitivity, or that perform error estimation. The direct consequence is that it currently remains unclear what is influencing the quality of home detection methods. The broader consequence is that many assumptions underlying home detection methods remain unproven or even uncovered. For example, one assumption often found in literature, is that the period and duration of observations do not influence the quality of home detection. Common sense suggest that this is not the case, as does research. Deville et al. (2014); **?** for example show how during summer months mobile phone users in France tend to move to touristic areas along the coast or near the mountains. Performing home detection on mobile phone data collected during this period will more likely lead to more wrongful results compared to using another period, proving one underlying assumption wrong. In our results too, we find this summer holiday effect, as can be observed in figure 3.

Our main argument is that such and similar sensitivities should be uncovered, acknowledged, and assessed when performing home detection methods as they will introduce errors and uncertainties that propagate in further steps of analysis.

Figure 3: Spatial patterns of the user counts obtained by the MA algorithm for a non-summer (A) and a summer (B) period of 14 day duration. Home detection in summer periods results in higher user counts in touristic areas, which in turn results in lower correlation with the ground truth dataset(C).



4 Acknowledgements

This work was funded by the EPSRC Center for Doctoral Training in Digital Civics [grant no. EP/L016176/1] at Open Lab, Newcastle University

5 Biography

Dr. Maarten Vanhoof is a Research Fellow at the Centre for Advanced Spatial Analysis in UCL. His PhD has focused on the use of mobile phone data for official statistics and geographical research.

Dr. Clement Lee is a Research Fellow at Open Lab in Newcastle University. His interests lie in the applications of spatial statistics to new and interesting datasets.

Dr. Zbigniew Smoreda is a senior researcher at Orange Labs, France. His work is focused on the use of mobile phone data to answer large-scale sociological questions

References

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., and Ratti, C. (2015). Choosing the Right Home Location Definition Method for the Given Dataset. In Liu, T.-Y., Scollon, C. N., and Zhu, W., editors, *7th International Conference on Social Informatics (SocInfo)*, pages 194–208, Beijing. Springer.

Cottineau, C. and Vanhoof, M. (2019). Mobile Phone Indicators and Their Relation to the Socioeconomic Organisation of Cities. *ISPRS International Journal of Geo-Information*, 8(1):19.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):15888–93.

Janzen, M., Vanhoof, M., Smoreda, Z., and Axhausen, K. W. (2018). Closer to the total? Long-distance travel of French mobile phone users. *Travel Behaviour and Society*, 11:31–42.

Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., and Giannotti, F. (2016). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics*, 2(1-2):75–92.

Vanhoof, M., Hendrickx, L., Puussaar, A., Verstraeten, G., Ploetz, T., and Smoreda, Z. (2017). Exploring the use of mobile phones during domestic tourism trips. *Netcom*, 31(3/4):335–372.

Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018a). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 3(4):935–960.

Vanhoof, M., Reis, F., Smoreda, Z., and Ploetz, T. (2018b). Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art. *Arxiv*, pages 1–13.