

Home location prediction with telecom data: benchmarking heuristics with a predictive modelling approach.

Dieter Oosterlinck^a, Philippe Baecke^b, Dries F. Benoit^a

^a*Faculty of Economics and Business Administration, Ghent University, Tweeckerkenstraat 2, B-9000 Ghent, Belgium*

^b*Area Marketing, Vlerick Business School, Reep 1, B-9000 Ghent, Belgium*

Abstract

Correctly identifying the home location is crucial for human mobility analysis with telecom data, more specifically call detail record (CDR) data. To that end, multiple heuristics have been developed in literature. Nevertheless, due to the lack of ground truth home location data, no study has thoroughly validated these widely used methods so far. We present a detailed performance analysis of existing home detection heuristics, using a unique dataset that enables this important validation on the lowest level, being the level of the cell tower. Our research indicates that simple heuristics surprisingly outperform their more complex counterparts. The benchmark study revealed that the best heuristic is able to identify the home location with an average error of approximately 4.5 kilometres and selects the correct home tower in 60.69% of the cases. Based on the insights provided by our study, we propose a new heuristic that increases the accuracy to 61% and lowers the average distance error to 4.365 kilometres. Secondly, if the home location is known for possibly only a fraction of the instances, we propose a labelled predictive modelling approach. Adding social network based variables in this predictive model further enhances the predictive performance. Our best model reduces the average distance error to 2.848 kilometres and selects the correct home location in 72.08% of the cases. Furthermore, this result provides an indication of the upper bound for home detection with CDR data. Finally, models that only make use of social network based data are developed as well. Results show that even without using data of the focal individual, these models are able to select the correct home tower in 37.65% of the cases and achieve an average distance error of 8.1 kilometres.

Keywords: Human mobility, Home detection, CDR data, Benchmarking, Predictive analytics, Social network analysis

Email addresses: dieter.oosterlinck@ugent.be (Dieter Oosterlinck),
philippe.baecke@vlerick.com (Philippe Baecke), dries.benoit@ugent.be (Dries F. Benoit)

1. Introduction

The number of studies about human mobility displayed a steep increase around 2008 and is still growing at a high pace (Barbosa et al., 2018). The interesting field of human mobility includes a large variety of applications and is therefore able to have a large impact on everyday life. Human mobility analysis sparks the development of smart cities, enables socio-economic studies, facilitates the understanding of mobility patterns and boosts better data-driven decision making amongst others. Vanhoof et al. (2018c) states that human mobility analysis will render important insights in the wider structures governing our society. The study of human mobility data is remarkably important for epidemiological studies in order to model the spread of viruses and even assess the impact of measures taken during the Covid-19 crisis. More applications can be found in the research about commute behaviour (Kung et al., 2014), commute distances, the impact of mobility on our carbon footprint (Isaacman et al., 2011) and even traffic prediction (Lv et al., 2014). Insights from human mobility analyses can further optimize telecom and transportation infrastructure.

Research has shown that telecom data obtained from mobile phone networks, call detail record (CDR) data, has great value for these analyses. Furthermore, Gonzalez et al. (2008) and Song et al. (2010) prove that human mobility is strongly predictable when using CDR data, as people spent most of their time in a limited number of locations. However, due to matters of confidentiality, CDR data typically lacks contextual information (e.g. the content of the messages or calls), which makes it not obvious to interpret the location traces in the raw data (Liu et al., 2013). It is therefore crucial to investigate methods that annotate the raw data into meaningful locations. The home location is one of the most important meaningful locations as the analysis of human mobility typically requires identifying the home location as a first step. This makes that home location prediction is very often a (first) part of more complex studies (Bojic et al., 2015). An accurate identification of the home location is therefore essential for this area of research. However, despite its large impact on the mobility analysis, literature failed to devote significant attention to this critical aspect. The main reason for this absence of attention has been the lack of proper validation, due to unavailability of ground truth data on an individual level. We fill this gap in literature by thoroughly evaluating the existing home detection methods using a unique data set that contains the closest cell phone tower to the actual home location. Throughout this study, the term *home tower* will be used to refer to this location. We extract the different categories of methods in literature and execute a benchmark study using 5 times 2-fold cross-validation in order to provide robust results.

The advantage of these heuristics is that they can be used even when no single home location is known. However, if the home tower is known for a part of the data set, we propose to use a labelled predictive modelling approach. Our results show that a labelled approach is able to significantly enhance the results. The performance of such a model also indicates to a large extent what the maximal attainable performance of home detection algorithms with CDR data is.

Previous research has already shown the value of social network data in multiple settings. In the context of home detection, research has been done in an online social network (OSN) (Backstrom et al., 2010). We will investigate the added value of using social network data in a CDR context. Furthermore, we build stand-alone social models that provide interesting insights for academics and telecom providers. The results show that using only the data of individuals in the social network of the focal individual, has predictive power for the home location of the latter, which also opens up possibilities for further research as it can be expected that this is valid for other than home locations as well. Telecom providers that are looking to increase their market share, might gain intelligence about non-customers by using this knowledge as well.

2. Literature review

2.1. CDR data for human mobility and the need for home detection

Research in the field of human mobility, urban planning, transportation engineering and mobility patterns was traditionally based on travel surveys, road side surveys and travel diaries (Calabrese et al., 2011; von Mörner, 2017; Wang et al., 2018). These survey methods have major shortcomings such as small sample rates, short survey durations, under-reporting and a high cost (Calabrese et al., 2011). Meanwhile, a variety of other data sources has been used, such as circulating bank notes (Brockmann et al., 2006), Foursquare check-in data (Noulas et al., 2012), tweets (Hawelka et al., 2014; Mahmud et al., 2014; Hironaka et al., 2016) and GPS data (Vazquez-Prokopec et al., 2013; Tang et al., 2015).

However, Barbosa et al. (2018) report in their review paper that call detail record (CDR) data is the most important, game-changing data of the last decade for analysing human mobility. CDR data is the information that telecom providers capture, every time that a customer makes or receives a call / SMS. Every record in a CDR data set contains interactional (a caller and receiver id), temporal (timestamp and duration) and, importantly, location aspects. The location refers to the geographical coordinates of the cell tower that is used and is therefore always an approximation of the actual location of the user. It was shown already by Gonzalez et al. (2008) and by Song et al. (2010) that human mobility is highly predictable when studied using CDR data. People will show different usage patterns on different key locations. From these it becomes possible to derive meaningful insights (Karikoski and Soikkeli, 2013; Blondel et al., 2015). An extensive review of the research on CDR data is published by Blondel et al. (2015).

CDR data is ideally suited for both large scale location analyses, as well as for research on individual level mobility. Mobile phones have a worldwide penetration rate of 96% (Iqbal et al., 2014). This makes that CDR data overcomes the low sample problem of the survey approach by capturing almost the entire population and offers a consistent approach for research on mobility patterns throughout the world (Kung et al., 2014). Using CDR data also implies tracking all mobile phones in a provider's network, not only the users that

installed a certain application on their device (Scherrer et al., 2018), which would be the case if GPS smartphone data is used. CDR data does also not require additional sensor data and map information, which substantially lowers the cost of data collection as well as being rather easily transferable to other regions (Liu et al., 2013). The problems of self-reported behaviour in surveys (Eagle et al., 2009) and their traditional high cost are also mitigated (Isaacman et al., 2011).

Of course, when using CDR data for location based applications, one needs to keep in mind that this data source was not originally designed for this objective. The original function of the data was to count the usage per customer for billing purposes. This makes that the observations are recorded only when somebody makes use of the network by calling or texting. The data generation is thus non-continuous. The precision level of the location also depends on the distribution of the cell towers. These aspects might lead to a low temporal and spatial granularity. One also needs to take into account the different market share of providers and the difference in calling plans between customers. However, most of these effects are largely mitigated due to the large geographic coverage and the high penetration of mobile phones (Wang et al., 2018).

Research has established the value of CDR data for human mobility analysis. Multiple authors (Kung et al., 2014; Vanhoof et al., 2018b; Bojic et al., 2015; Dash et al., 2014) indicate that the correct identification of the home location is an important prerequisite for the large majority of applications in human mobility, such as home-work commuting, commuting patterns, mobility profiles, mobility and epidemiological models (Tizzoni et al., 2014). Isaacman et al. (2011) point out that people spend most of their time on a limited number of locations. It is crucial to identify these key locations, such as the home location, in order to understand human mobility, social patterns and implement technology and policy decisions like the deployment of telecommunications and transportation infrastructure. Nevertheless, the methods of home detection are often obscured in literature (Vanhoof et al., 2018b). Despite that this initial step largely determines the quality of the subsequent analyses, the performance is hardly ever validated. The absence of a ground truth is in many cases the main reason why this validation can not be done. Vanhoof et al. (2018b) underlines that research with individual level ground truth needs to be done in order to assess the quality of the different home detection methods.

2.2. Categories of home detection algorithms

The home location in CDR data usually refers to the coordinates of the *home tower*, the cell tower that is closest to the actual home location. As the level of precision is restricted to the level of the cell towers, this is also the case for the home locations and thus *home towers*.

Literature commonly distinguishes two broad classes of home detection methods, as we report in Figure 1. *Single-step* methods apply home detection rules directly on the individual towers. The second category, the *two-step* methods, add an extra initial step by first clustering towers. This is done in order to counter the fact that cell phones might

switch towers despite remaining at the same location. In the second step, these methods apply home detection rules similar to the single-step methods, to these clusters. Hence, both approaches need decision rules in order to identify important places and label the correct location as home (Vanhoof et al., 2018b). Based on the literature, we further divide the single-step methods into two classes, the activity and inactivity heuristics. The former takes the standard approach by using data when somebody is actively using his or her phone, whereas the latter models the periods in between usage. This results into three classes, presented in Figure 1.

2.2.1. Decision rules and heuristics based on activity (Activity heuristics)

This first category of home detection algorithms contains single-step methods. Decision rules are investigated to identify the home tower. Vanhoof et al. (2018b) examine three decision rules that fit in this category. A first heuristic, also used by many other authors (e.g. Tizzoni et al. (2014); Song et al. (2010)), states that the home tower is the cell tower where the majority of both outgoing and incoming calls and texts are observed, also called the amount of activities criterion, we will refer to this method as Act_1.

A second decision rule aims to improve upon the former heuristic by taking into account the regularity of a certain location. An extraordinary event might lead to a lot of exceptional calls made or received by someone on a certain location. Applying the Act_1 method to this case, would result into selecting this exceptional location as most used and the location will therefore be labelled as the home location. The aim of the second criterion is to counter this undesirable effect. Regularity will be modelled with the amount of distinct days criterion. If somebody is regularly on a certain place, the chances are higher that this is the home location. The Act_2 method counts the amount of distinct days on which a location was used. The tower with the maximum number of distinct days with phone activities is then selected as home location.

Another attempt at further improving the amounts of activities criterion (Act_1) is done by including time constraints. These time constraints heuristics add the restriction that only activities between specific hours should be taken into account. The aim is to select a time frame during which people are most likely to be at home. Vanhoof et al. (2018b) select the time frame from 7PM till 9AM (Act_3). By selecting such time frame, the observations during working hours are excluded. The underlying assumption is that people are less likely to be at home during working hours and more likely to be at home during non-working hours. Other authors (Phithakkitnukoon et al., 2012; Phithakkitnukoon and Smoreda, 2016; Bachir et al., 2019) use this same criterion, but with a different parameter choice. They use a more stringent range and try to model night time, for which the time frame of 10PM until 7AM is selected (Act_4).

A final activity based heuristic combines the ideas in the above methods. Calabrese et al. (2011) apply both the concept of regularity and the refinement in terms of time frame. This results into the number of distinct nights criterion. The authors select the time frame from 6PM until 8AM as night time. The number of nights with activity on

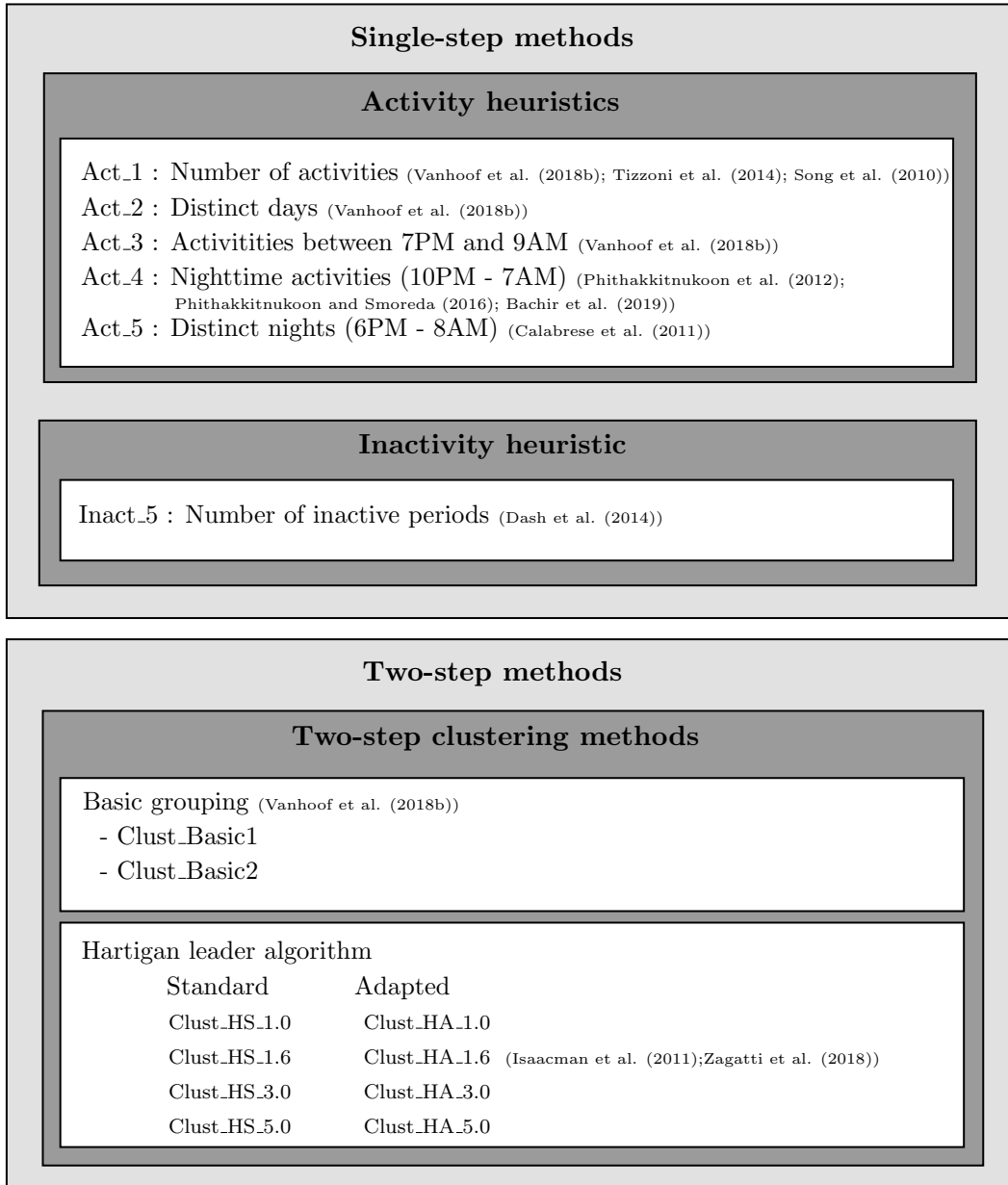


Figure 1: *Overview home detection methods.* Three main categories of home detection methods were identified in literature; activity heuristics, inactivity heuristics and two-step clustering methods. The different methods that will be evaluated in our benchmark study can be found in this table.

each location is counted and the location with the maximum number of distinct nights is selected as home location (Act_5).

It is clear that for the time restricted methods, the parameter choice will influence the performance. However, since proper validation with individual level ground truth lacks in literature, it has been difficult to select the optimal parameters. We will therefore benchmark these methods on our unique dataset.

2.2.2. Decision rules and heuristics based on inactivity (Inactivity heuristic)

Instead of taking into account the activity of a user on a certain location, this second category counts the number of times that somebody is inactive on that location (Dash et al., 2014). Given that one usually sleeps at his/her home location, the idea behind the inactivity heuristic is to set a threshold for a period without activities in order to model the sleeping hours. In practice, this means that a location is counted if the time between an activity on that location and the next activity (on any location) is longer than the selected threshold. In other words, the periods with no data are used and the last location preceding this empty period is registered. The advantage is that this also works for shift workers, as inactivity does not need to be observed during specific hours during night time. The location with the highest count of inactivity is selected as the home location. The threshold was set to five hours by Dash et al. (2014).

2.2.3. Two-step clustering approaches

The two-step, clustering, approaches deviate from the single-step approaches by first clustering certain towers together. Because of physical boundaries and other properties of the cell towers and the environment, it is possible that the cell phone switches connection between different towers, although the user stays at the same location. In order to mitigate this effect, towers are clustered together based on their location. This first step is executed for every individual, thereby resulting into different clusters of towers for every individual. The second step is to label the clusters by scoring them according to criteria, similar to the single-step methods.

The idea of clustering certain towers is closely related to the fundamental concept of stemming in the text clustering literature (Porter et al., 1980; Bharti and Singh, 2015; Abualigah et al., 2018a,b; Abualigah, 2019). The application of clustering as a first, pre-processing, step in the two-step clustering approaches, aims to avoid biases by clustering similar elements together. In text clustering the elements are words, in our application the elements refer to cell towers. Stemming transforms inflectional forms of certain words to the same root or *stem* by removing the prefixes and suffixes of each word (Abualigah et al., 2018b). For example, the words consult, consultant, consulting and consultative will result in the same stem ‘consult’. It is evident that for many applications the difference between the individual words is irrelevant and that the performance of the resulting model will significantly improve due to stemming. The clustering of cell towers aims to achieve a similar improvement, by clustering certain towers based on geographical proximity, thereby reducing detrimental differences in location.

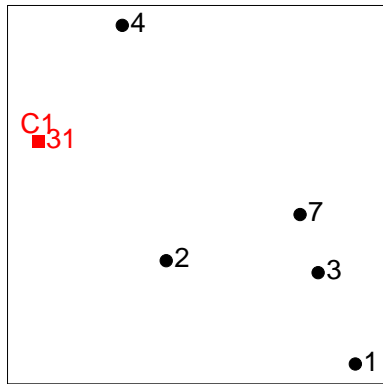
Different options for the first, clustering, step are examined in literature. Vanhoof et al. (2018b) present a first example of clustering cell towers. This basic approach clusters all activities for a selected individual that are recorded within a spatial perimeter of 1 kilometre around the cell tower. Thus, if we look at one cell tower, all activities of the user on other towers within the 1 kilometre perimeter are added to the first tower. This approach differs from the other clustering approaches by not scoring the clusters as a whole, but scoring every tower by including the records of the towers within their cluster. This results into scores for every tower, as opposed to scores only for every cluster. The scoring for the basic clustering by Vanhoof et al. (2018b) is selected from their activity heuristics. Vanhoof et al. (2018b) apply their first activity heuristic (Act_1) resulting into the first option; Clust_Basic1, where the home tower will be selected as the tower with the highest number of activities on this tower, including the activity within the perimeter. They also use their third activity heuristic (Act_3) resulting into Clust_Basic2, which is identical to Clust_Basic1, except that only activities between 7PM and 9AM are taken into account.

Other clustering approaches are presented by Isaacman et al. (2011) and Zagatti et al. (2018). In the first step of this two-step procedure, cell towers are clustered based on the Hartigan leader algorithm (Hartigan, 1975). One of the advantages of this clustering algorithm is that it does not require an a priori chosen number of clusters. We will discuss both the standard algorithm and an adapted version which was used by Isaacman et al. (2011) and Zagatti et al. (2018).

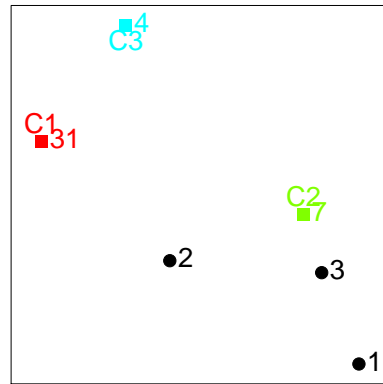
The standard Hartigan leader algorithm starts by ranking the observations (in this case, the cell towers) based on a selected feature (in this case the number of distinct days). The algorithm assigns the first tower in the list as the cluster centre for a first cluster. The algorithm then descends through the sorted list of towers and checks whether the next tower falls within a chosen threshold distance from an existing cluster centre. If this is the case, the tower is added to the existing cluster, if not, a new cluster is formed with this tower as centre.

Isaacman et al. (2011) and Zagatti et al. (2018) implemented the algorithm with the adaptation that the cluster centres can move. If a new tower is added to an existing cluster, the cluster centre is recalculated, weighted by the distinct days that a tower is used. This process can be observed in Figure 2.

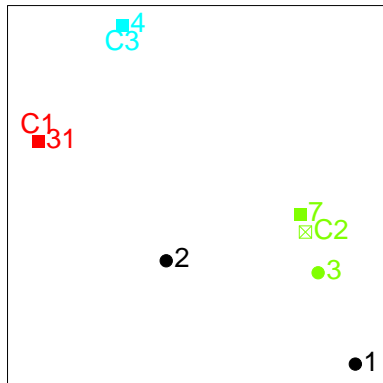
In the second step of this two-step procedure, the clusters need to be scored. Zagatti et al. (2018) propose to score the clusters by summing +1 for every activity in the evening (7PM - 7 AM) and during the weekend and -1 for daytime hours (8AM-5PM) and weekdays (Scoring a in Figure 1). Furthermore, the algorithms can be executed with different thresholds for the spatial perimeter. Isaacman et al. (2011) and Zagatti et al. (2018) use 1.6 kilometres, but discuss that other options could be suitable as well. This study will investigate threshold values ranging from 1 to 5 kilometres. The naming convention for the different implementations of this category in Figure 1 is as follows. HS refers to the standard Hartigan leader algorithm, while HA refers to the adapted version. The number refers to the threshold (in kilometres) that is set in the algorithm.



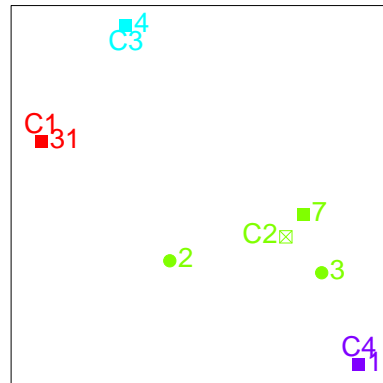
(a) Selection of first cluster centre (C1).



(b) Allocation of subsequent cluster centres (C2, C3).



(c) Weighted re-calculation of cluster centre C2.



(d) Final clusters.

Figure 2: *Example of adapted Hartigan leader algorithm.* The algorithm starts with the calculation of the number of distinct days (cf Act.2) for every tower, represented by the numbers in the figure. The tower with the highest number is selected as the first cluster centre, indicated with a square (a). The algorithm then selects the tower with the second highest number of distinct days and so on. If none of the distances to an existing cluster centre satisfies the threshold value, the tower is selected as a new cluster centre, in this example resulting into the clusters with centre C2 and C3 (b). The distance from the fourth tower in the ranked list (with 3 distinct days) to cluster centre C2 satisfies the threshold. This is the first tower in this procedure that can be added to an existing cluster. The cluster centre needs to be re-calculated due to the addition of the new tower. The new centre is weighted by the number of distinct days, explaining why C2 now lies closer to the original cluster centre; the tower with 7 distinct days (c). The square with an x inside indicates that the cluster centre is the result of the weighted re-calculation. The algorithm continues by adding the observation with 2 distinct days to the second cluster, which leads to a new cluster centre again. The last tower does not satisfy the threshold to any existing cluster centre and results into a final single-observation cluster C4 (d).

2.3. Validation

As mentioned earlier, one of the main shortcomings in literature is the lack of proper validation. Validation is either not done (e.g. Zagatti et al. (2018)) or limited to validation on an aggregated level, usually through census data (e.g. Phithakkitnukoon et al. (2012); Kung et al. (2014); Ahas et al. (2010); Calabrese et al. (2011)). This means that the sum of estimated home locations is compared to the population count at that location. Common validation metrics on the aggregated level measure the correlation between the count of estimated home locations and the census data. Examples are the cosine similarity metric (Vanhoof et al., 2018b), Pearson’s r (Vanhoof et al., 2018a) or a simple matching coefficient (Bojic et al., 2015). Two studies contain a small sample ground truth and report distance between the estimated and actual location (Isaacman et al., 2011; Dash et al., 2014).

The problem is that the level of granularity of the census data (e.g. city or village level) is very unlikely to align with the granularity level of the CDR data (the distribution of the cell towers). An additional issue is that this type of validation gives no guarantee whether the algorithm actually detects the correct individual home location as illustrated by the following simplified example. Census data reports that both city A and city B count 1000 inhabitants. Now, consider an algorithm that would actually locate the entire population of city A in city B and vice versa. It is clear that the accuracy should be zero, however as validation is only done on the aggregated, city, level, the reported accuracy will be 100%. Third, using CDR data is generally limited to one provider. The spatial distribution of the market share of this one provider is typically unknown, which also hinders this type of validation. The fact that there is no consensus in literature on which home detection methods are best, is to a large extent due to these validation problems.

It is clear that there is a high need for a thorough validation on the level of the individual, as already suggested by Vanhoof et al. (2018b). The anonymised CDR data in this study contains the tower closest to the home location of the individual users and therefore enables the crucial validation.

2.4. Social network

In their review paper about human mobility research, Barbosa et al. (2018) discuss that individuals in a social network, such as friends, family or colleagues are likely to share locations and mobility patterns (Axhausen, 2005; Carrasco and Miller, 2006; Dugundji and Walker, 2005). In the context of online social networks, Liben-Nowell et al. (2005) were among the first to show a relation between distance and online friendships. Backstrom et al. (2010) find that the location of close contacts may predict the location of the focal individual in the online social network (OSN). Bojic et al. (2015) confirm that it is possible to identify meaningful places such as home location, for users that do not reveal any location information themselves, by using merely the location data of their friends on the OSN. Backstrom et al. (2010) report that the home location can be estimated within 40 kilometres of their actual home location for almost 70% of the US-based Facebook users that have more than 15 friends.

Multiple researchers confirm these findings in the context of mobile phone users (Wang et al., 2011; Lambiotte et al., 2008; Krings et al., 2009; Phithakkitnukoon et al., 2012). Furthermore, Phithakkitnukoon et al. (2012) discuss that a mobile phone social network derived from CDR data is a better representation of actual everyday personal networks than online social networks. Both the insight that the location of other individuals in the social network has predictive power for the home location of the focal individual and the insight that CDR based social networks outperform online social networks, are a clear indication that augmenting the home estimation models with social network data might add to the performance of the models.

The results of Backstrom et al. (2010) in an online social media context, can even be enhanced by adding information about how strongly individuals are connected (Chen et al., 2014). Phithakkitnukoon et al. (2012) showed that around 80% of visited locations are within 20 kilometres of peoples nearest social ties locations. Increasing this geo-social radius to 45 kilometres makes the figure rise to 90%. Hence, it is valuable to take the concept of *tie strength* into account to further enhance the predictive performance in the home location prediction models as well. We will define tie strength based on previous research (Onnela et al., 2007; Nitzan and Libai, 2011; Roelens et al., 2016; Meyners et al., 2017). The tie strength between individual i and individual j is defined as the ratio of the communication volume between individual i and individual j and the total communication volume of the individual i . As suggested by Nitzan and Libai (2011) and later used by Roelens et al. (2016), the weight of a text message is set equivalent to a one minute call.

$$Comm_{ij} = sms_in_{ij} + sms_out_{ij} + minutes_calls_in_{ij} + minutes_calls_out_{ij} \quad (1)$$

$$Comm_i = \sum_j sms_in_{ij} + sms_out_{ij} + minutes_calls_in_{ij} + minutes_calls_out_{ij} \quad (2)$$

$$Tiestrength_{ij} = \frac{Comm_{ij}}{Comm_i} \quad (3)$$

3. Methodology

The methodology section introduces the four important methodological aspects of this research. First, we introduce the dataset that will be used for the main parts of the analysis. Secondly, we explain our validation procedure, crucial to this research. Thirdly, we shortly discuss how we benchmark the heuristic methods encountered in literature to our dataset, including some value adding adaptations. In a fourth methodological section, we introduce our predictive modelling approach for the problem of home detection.

3.1. Data

The CDR data used in this study contains five weeks of anonymised CDR data including voice calls and SMS. For an effective evaluation of home detection methods, it is advised to avoid holiday periods as too many people display a temporary change in behaviour and might even be away from home for several weeks (Blondel et al., 2015). Data of the provider in the period from Monday 2 May 2016 until Sunday 5 June 2016 was selected for this purpose.

The ground truth home location in our data set is based on the billing address of the customer. As discussed in Section 2.2, this billing address is substituted by the cell tower that is closest to this location and will be referred to as *home tower*. One needs to be aware that people may divide their time between more than one home location. People might have a secondary home in which they spend time during weekends for example. However, in line with previous research, we aim at modelling the main home location. Although our ground truth can not provide conclusive evidence that we are in fact dealing with the main home location for every single individual, we can assume that this is the case for the vast majority of the data set. As people generally receive their bills on their main home address, using the billing address for this purpose can be considered as a suitable choice for defining the ground truth home location.

Several researchers remove less active customers, by requiring a minimum threshold number of used towers (Barbosa et al., 2018), a minimum number of calls (Kung et al., 2014; Dash et al., 2014; Bojic et al., 2015) or a minimum number of days with observations (Ahas et al., 2010). This filtering however limits the scope and artificially boosts the performance as only cases with a lot of data points are retained. It is of course much easier for the algorithm to detect the correct home location for these more informative cases. In other words, a too strict filtering artificially improves the performance. We opted for a very light restriction of minimum 5 observations in the five week period, so that we at least exclude idle numbers. Our intention is to compare the different methods on an equal, fair basis and to make sure that the results have a broad scope, including less active users. We retain 93.57% of the users by imposing this mild threshold.

A random sample of 100,000 customers satisfying the threshold is selected. This results into 54,567,294 records or 15.59 observations per day for the average customer. 100,000 customers results into 100,000 home towers in the data set. Customers also use non-home towers, this leads to 2,159,444 tower - customer id combinations. The average customer thus used 22.59 towers, in other words, on average, 4.43% of the used towers are home towers.

3.2. Validation metrics

Whereas literature assessed the home detection methods merely on an aggregated level, we are able to assess the performance on individual level. We will calculate two important measures at this level: accuracy and distance error.

Accuracy is calculated as the percentage of predicted home locations (home towers) that are actual home towers. It is important to keep in mind that a random model would achieve an accuracy of 4.43% as this is the percentage of home towers in the data.

As the accuracy alone does not capture the entire picture, the average distance between the predicted and the actual home tower will be used as a second validation metric. Consider an algorithm that predicts a location as home, that is 100 kilometres separated from the actual home tower. It is logical that this case will be counted as incorrect in the accuracy measure. However, if the same algorithm predicts a location only 1 kilometres separated from the home tower as home, it will also be counted as incorrect, although the algorithm performs very well in this case. The accuracy measure might therefore be a serious underestimation of the actual performance, as several towers might be an acceptable solution. The distance measure takes this into account.

Furthermore, for robustness of the results, we use five times two-fold cross-validation (5x2cv) (Alpaydin, 1999). In a first step, the 5x2cv method randomly splits the data into two folds. Each fold is used once as a training set and once as test set. This procedure is repeated five times. As opposed to predictive modelling method that we present in Section 3.4, the benchmark heuristics in the following section do not require a training phase. Performance of these methods can therefore immediately be assessed on the test sets resulting from the 5x2cv. For both approaches, this leads to ten test values for every metric. The 5x2cv method also encompasses the 5x2cv F-test to assess the significance of the difference in performance between the models (Alpaydin, 1999).

3.3. Benchmarks

We deploy the methods described in literature (Section 2.2) to our CDR data set described in Section 3.1. For the two-step clustering algorithm, next to the proposed 1.6 kilometres of Isaacman et al. (2011) and Zagatti et al. (2018), we also investigate a threshold of 1, 3 and 5 kilometres. Furthermore, whereas single-step methods identify one tower as home location, the two-step methods can result into a new cluster centre that lies in between towers. For validation purposes, the identified cluster centre needs to be brought back to the closest tower to this centre. This makes that the results are evaluated on the correct level of granularity and that they are compared on a fair basis between the different categories of home detection algorithms.

3.4. Predictive modelling approach

The presence of a home location in our unique dataset not only allows for a thorough validation of existing measures, but also allows for a labelled predictive modelling approach. We can use the labelled data to train models that learn how to identify the home location. To our knowledge, the only research that adopted a similar approach was done by Liu et al. (2013). Their dataset was however restricted to only 80 users with labelled data. Furthermore, their study was not aimed specifically at home detection.

We will first explain how the models are constructed, with respect to the decisions about the dependent variable, the independent variables and the classification algorithms. It is to be expected that using state-of-the-art classification algorithms on a labelled dataset will increase the home detection performance. Of course, the scope of a labelled approach is more limited than the unlabelled heuristic methods, as this method is only applicable if at least some home locations are known. This method also provides a substantiated indication of the maximal attainable performance of home detection algorithms based on CDR data.

3.4.1. Binary dependent variable home tower

In order to build a model that will identify the most likely home tower, we need to feed our algorithm with observations to learn from. We construct a base table that contains observations for home towers as well as non-home towers. Both categories are needed for the model in order to learn how to distinguish between both. The structure of the base table is represented in Table 1.

Selecting observations for the home towers (class 1) is straightforward, as we know for each user in our training set what the home tower is. Selecting non-home towers is more involved as this choice is less obvious and will affect the results. In theory, every tower that is not the home tower for a certain user can be selected as part of this class (class 0). However, this would mean that there would be zero activity for the selected user on the majority of the towers in this class. The resulting model would only learn to distinguish between used and non-used towers and will therefore be useless for the identification of the home location. We therefore select only towers that have been used at least once. By doing this, it becomes much harder for the model to distinguish between both classes. However, the model will be much more informative and relevant.

Id	Tower	Dependent variable	Independent variables		
			calls_in_nbr	calls_in_dur_total	...
1	home tower	1	5	210	...
1	non-home tower 1	0	2	103	...
1	non-home tower 2	0	1	504	...
2	home tower	1	10	1243	...
2	non-home tower 1	0	3	239	...
2	non-home tower 2	0	12	2087	...
2	non-home tower 3	0	2	96	...
...	...	1/0

Table 1: *Structure of the base table for the predictive modelling approach.* The dependent variable indicates whether the tower is the home tower for the id. The independent variables are calculated based on the CDR data observed on that tower for the id. This structure results into a base table with home and non-home towers and can therefore be used to build a model that distinguishes between both types of towers. The tower with the highest predicted home tower probability will be selected as the home location.

3.4.2. Independent variables

We constructed 30 independent, or explanatory, variables. Twenty-two of these are constructed based on the three categories identified in Section 3.3. Eight social network based variables are included as well. We present the variables in Table 2. Note that it follows from the structure of the base table that every variable is calculated per user, per tower.

In order to use the logic in the home detection heuristics, the variables needed to be translated in order to fit in the structure of the base table. The Act_2 heuristic for example selects the tower that had the most distinct use days as the home tower for the selected user. Translating this into an independent variable, this becomes the number of distinct days on this tower, for this user. It is straightforward to see how the other heuristics were translated into variables as well, following the same logic.

The first activity heuristic, Act_1, selects the tower with the highest number of activities. We decided to split this into multiple variables, separated into incoming versus outgoing and voice calls versus text messages. We also enriched this by calculating other measures such as average and standard duration of calls and the percentage of the activity of the user on this tower.

Literature showed the predictive power of social networks. We therefore augmented the base table with variables that take into account this social network. We included three variables based on how frequently the contacts of an individual also use a certain cell tower. Five variables were included based on tie strength as defined in Section 2.4.

Furthermore, we will explore the added value of social network data by building three categories of models: the *full* models are trained on all variables, the *withoutsocial* models do not use any social network based variables, whereas the *socialonly* models do only use the social network based variables.

3.4.3. Binary classification algorithms

We will examine four frequently used binary classification algorithms; logistic regression, random forest, adaboosting and neural network models. R statistical software was used to implement these models (R Core Team, 2020). The random forest models were run with 1,000 trees, as recommended by Breiman (2001). The adaboosting models are implemented following the method of Friedman et al. (2000) and allowed 150 boosting iterations. The neural network models are implemented with 40 units in the hidden layer and we restricted the algorithm to perform a maximum of 2,000 iterations. Evaluating multiple classifiers assesses the robustness of the labelled predictive modelling approach.

The output of these models is the probability of belonging to a certain class. For every user, the tower with the highest predicted home tower probability is selected as the home tower. We will evaluate the predictive modelling approach based on the same measures as the benchmark methods; accuracy (percentage correctly predicted home towers) and the average error distance to the actual home tower, thereby following the same 5x2cv procedure.

Activity based	calls_in_nbr	Number of incoming calls.
	calls_in_dur_total	Total duration of incoming calls.
	calls_in_dur_avg	Average duration of incoming calls.
	calls_in_dur_sd	Standard deviation of duration of incoming calls.
	calls_in_perc_on_tower	Percentage of incoming calls.
	calls_out_nbr	Number of outgoing calls.
	calls_out_dur_total	Total duration of outgoing calls.
	calls_out_dur_avg	Average duration of outgoing calls.
	calls_out_dur_sd	Standard duration of outgoing calls.
	calls_out_perc_on_tower	Percentage of outgoing calls.
	sms_in_nbr	Number of incoming text messages.
	sms_in_perc_on_tower	Percentage of incoming text messages.
	sms_out_nbr	Number of outgoing text messages.
	sms_out_perc_on_tower	Percentage of outgoing text messages.
	act_2_distinct_days	Number of distinct days.
	act_3_7pm_9am	Number of activities between 7PM and 9AM.
	act_4_nighttime_10pm_7am	Number of activities between 10PM and 7AM.
	act_5_distinct_nights_6pm_8am	Number of distinct nights (6PM - 8AM).
Inactivity based	inact_5	Number of inactive periods (>5h).
	inact_7	Number of inactive periods (>7h).
Clustering based	clust_Basic1	Number of activities within a 1 kilometre perimeter.
	clust_Basic2_7pm_9am	Number of activities, 1 km perimeter (7PM - 9AM).
Social network	soc_sum_cdr	Sum of number of activities of contacts.
	soc_nbr_contacts_use_loc	Number of contacts that use this location.
	soc_perc_contacts	Percentage of contacts that use this location.
	soc_tiestrength_avg	Average tie strength with contacts on this location.
	soc_tiestrength_median	Median tie strength with contacts on this location.
	soc_tiestrength_min	Minimum tie strength with contacts on this location.
	soc_tiestrength_max	Maximum tie strength with contacts on this location.
	soc_tiestrength_sum	Sum of tie strength with contacts on this location.

Table 2: *Variables in predictive model.* Every variable is calculated based on the structure of the base table, therefore ‘on this tower for this id’ can be added to the description of every variable.

4. Results

4.1. Benchmarks

We present the results of our benchmark study of the heuristic home detection methods introduced in Section 2.2 in Table 3. These methods do not need a training phase, as opposed to a predictive modelling approach and are therefore immediately applied to the test folds in the 5x2cv approach. The reported numbers are the averages of the metrics over the 10 test folds.

Model		Correct home tower (%)	Distance to home tower (km)
Activity	Act_1	57.56	5.671
	Act_2	60.16	4.591
	Act_3	56.54	5.800
	Act_4	41.69	8.425
	Act_5	59.11	4.816
Inactivity	Inact_5	60.69	4.499
Two-step clustering	Clust_Basic1	53.69	6.885
	Clust_Basic2	55.01	6.276
	Clust_ha_1	6.41	25.133
	Clust_ha_1_6	6.03	26.155
	Clust_ha_3	5.34	28.850
	Clust_ha_5	4.87	32.587
	Clust_hs_1	5.88	25.517
	Clust_hs_1_6	5.15	26.875
	Clust_hs_3	3.76	30.456
	Clust_hs_5	3.22	34.898

Table 3: *Benchmark results.* The results indicate that the inactivity category performs best, with the highest accuracy (% correct home tower) and the lowest average error distance, followed by the activity and two-step clustering based methods. The maximal attainable performance with the existing home detection algorithms lies in the range of 60% accuracy and an average error distance of 4.5km.

The best performing category is the inactivity category. The average accuracy is 60.69% and the average distance error is only 4,499 metres. The Inact_5 method scores better than the best activity method (Act_2). The difference is highly significant with an F-value of 72.44 and an associated p-value of 8.95e-05 for the 5x2cv F-test for difference in accuracy. The idea of modelling sleeping hours, that is the basis of this heuristic, does not require the choice of a specific time of day and therefore allows for a correct home identification for a much larger range of people with different behaviour. This makes this heuristic much more broadly applicable, across cultures and countries as well. Despite its simplicity, the inactivity method also scores a lot better than the more complex two-step clustering methods.

The performance within the second best category, the activity category, is rather consistent. The best method is Act_2, which states that the home tower is the tower that is used on the maximum number of distinct days. The underlying assumption of regularity in this method seems to improve the performance a lot, when compared to Act_1 for example. This method selects the correct home tower in 60.16% of the cases, while the average error in terms of distance is limited to 4,591 metres. Act_2 is significantly better than the second best activity method (Act_5) (F-value 52.89, p-value 1.9e-3).

A surprising result is that the, more complex, clustering methods do not improve the performance. Comparing the two basic clustering methods with their single-step counterparts (Act_1 for Clust_Basic1 and Act_3 for Clust_Basic2), confirm this. The Hartigan leader based approach tremendously reduces performance, the best model using this approach achieves an accuracy of only 6.41%, whereas even the worst non Hartigan leader based algorithm still achieves 41.69%. In terms of distance, a similar performance drop is observed. We also observe that the adapted version (HA) performs slightly better than the standard version (HS). The higher the threshold value, the further the method deviates conceptually from the single-step methods. The results indicate that a higher threshold value further reduces performance. In the next section of the paper, more settings for the Hartigan based methods will be explored in order to identify what causes their unacceptable performance.

Furthermore, it is remarkable that the heuristics that take into account specific hours for night time (Act_4, Act_5 and Clust_Basic2) do not achieve a higher performance than their counterparts that do not take time into account. The lower performance can partly be explained due to the fact that part of the users do not have any data points during the specified time frame. This obviously limits the maximal performance that can be achieved. Table 4 reports the percentage of users for which no prediction can be made because of this issue.

Model	No Prediction
Act_3	1,67%
Act_4	16,12%
Act_5	0,15%
Inact_5	0,13%
Inact_7	0,14%
Clust_Basic2	1,67%

Table 4: *Percentage of individuals for which no prediction can be made.* Due to imposing a time frame in model Act_3, Act_4, Act_5 and Clust_Basic2, it is impossible to make a home location prediction for individuals that have no observations during the selected time frame. Act_4 uses the most restricted time frame, which results into the highest number of these cases. The inactivity methods will produce no prediction if no inactivity period is observed.

4.2. Optimisation of benchmarks

Given that the inactivity method performed best, we decided to further investigate this method. The question is whether optimising its main parameter, being the threshold value for the number of idle hours, further enhances the results. We present the results in Table 5 and graphically in Figure 3, where Inact_X means the inactivity method with a parameter value of X. Based on the accuracy metric, Inact_6 can be selected as optimal. However, based on the distance metric Inact_7 is optimal. The negligible difference in accuracy between Inact_7 and Inact_6 however is not significant (F-value 0.55, p-value 0.803). We therefore suggest to use Inact_7. Furthermore, the difference between this optimal setting and the arbitrarily chosen value of 5 by Dash et al. (2014) is significant (F-value 6.78, p-value 0.024). We also observe that the performance parameters show an inverted U shape. The performance increases gradually when starting with a threshold of 3 hours, to reach its maximum at the 6/7 hour threshold, after which the performance steadily declines again. This optimisation further demonstrates the added value of having high quality validation data.

Model	Correct home tower (%)	Distance to home tower (km)
Inact_3	60.05	4.748
Inact_4	60.48	4.612
Inact_5 (Dash et al., 2014)	60.69	4.499
Inact_6	60.81	4.455
Inact_7	60.80	4.413
Inact_8	60.73	4.421
Inact_9	60.61	4.448
Inact_10	60.33	4.556
Inact_11	59.90	4.669

Table 5: *Inactivity method.* Optimising the parameter of the inactivity function reveals that the arbitrary 5-hour value of Dash et al. (2014) does not lead to optimal performance. The Inact_7 method is selected as optimal, as it reaches the lowest distance error and an accuracy that is not significantly different from the optimal accuracy of the Inact_6 method.

The initial analysis revealed that the Hartigan leader based two-step methods have a much lower performance than the other methods. In order to further investigate this and better benchmark this category with the single-step methods, we decided to use different scoring rules in the second step. We score the clusters with the Act_2 and the Inact_5 method. The results are reported in Table 6.

First of all, we again observe that the adapted Hartigan leader algorithm (HA) performs better than the standard version (HS). Also, except for the combination of the adapted Hartigan leader algorithm with Inact_5, a higher threshold distance leads again to lower performance, indicating again that the further the method deviates from the single-step methods, the worse.

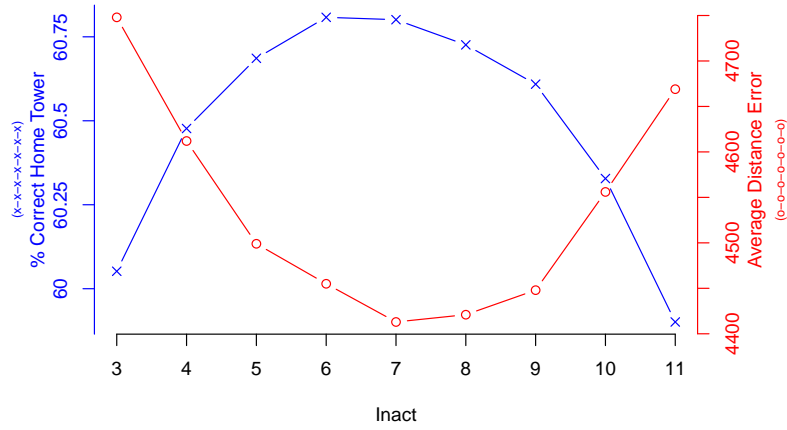


Figure 3: *Inactivity method*. The optimisation exercise shows a continuous shape for the performance functions. The value of 5 as suggested by Dash et al. (2014) does not lead to optimal performance. In terms of average distance error, 7 is the optimal value. In terms of percent correct home tower, 6 is optimal, however the difference with 7 is not significant. We can therefore conclude that 7 is the best value.

Clust_HA_1_Act_2 is now the best two-step counterpart for the single-step Act_2 method. The accuracy however drops from 60.16% to 55.70% and the distance error rises from 4.591 to 5.167 metres. This difference is highly significant (F-value 1248.96, p-value 7.51e-08). The same holds for the Inact_5 versus the Clust_HA_1_Inact_5 method; from 60.69% to 58.28% and from 4.499 to 4.525.

Model	Correct home tower (%)	Distance to home tower (km)
Clust_ha_1_act_2	55.70	5.167
Clust_ha_1_6_act_2	49.40	5.534
Clust_ha_3_act_2	37.82	5.700
Clust_ha_5_act_2	30.04	5.582
Clust_hs_1_act_2	46.75	5.489
Clust_hs_1_6_act_2	36.53	5.883
Clust_hs_3_act_2	20.47	6.150
Clust_hs_5_act_2	14.71	6.259
Clust_ha_1_inact5	58.28	4.525
Clust_ha_1_6_inact5	53.33	4.532
Clust_ha_3_inact5	41.77	4.493
Clust_ha_5_inact5	32.35	4.490
Clust_hs_1_inact5	50.12	4.600
Clust_hs_1_6_inact5	40.59	4.686
Clust_hs_3_inact5	22.62	4.983
Clust_hs_5_inact5	15.39	5.373

Table 6: *Scoring the Hartigan leader based methods with known activity/inactivity heuristics.* The scoring method proposed by Zagatti et al. (2018) resulted in bad performance. We therefore executed the clustering method with scoring methods based on the best inactivity/activity heuristics. This enhances the results for the clustering method strongly. However, when compared to the single-step methods, the performance is reduced despite the higher complexity of the two-step method.

Our results indicate that the most important part of the two-step approach is the second step, as the performance of scoring with better measures (Act_2 and Inact_5) clearly outperform the standard option. This second step corresponds to the single-step methods and underlines the importance of a well chosen heuristic. Nevertheless, this study shows that adding a clustering step to the single-step methods does not add any value.

The performance depends heavily on the assumptions and parameters in the heuristic. An optimal approach would be able to automatically select the best parametrisation and assumptions in every case. This idea is more embedded in a pure predictive modelling or classification approach.

4.3. Predictive modelling approach

The results of this approach can be found in Table 7. We will first compare these results with the benchmark heuristic approach, followed by a discussion of the different binary classifiers and a discussion of the added value of the social network data.

Model	Correct home tower (%)	Distance to home tower (km)
full_logreg	61.36	4.453
full_rf	71.86	2.952
full_adaboost	71.86	2.925
full_neuralnet	72.08	2.848
withoutsocial_logreg	60.78	4.682
withoutsocial_rf	71.42	3.094
withoutsocial_adaboost	71.40	3.070
withoutsocial_neuralnet	71.66	2.941
socialonly_logreg	37.65	8.098
socialonly_rf	32.74	9.712
socialonly_adaboost	37.64	8.480
socialonly_neuralnet	38.15	8.365

Table 7: *Predictive method results.* The improvement of the logistic model, when compared to the best unlabeled heuristic methods is minor. However, the other classifiers strongly enhance the results. The best method (full_neuralnet) reduces the average distance error with 1,565 metres and improves the correct home percentage with 11.28 percentage points.

The best predictive model is the neural network model that uses all created variables (*full_neuralnet*). This model predicts the correct home tower in 72.08% of the cases and has an average distance error of only 2.848 metres. Compared to the results of the optimal heuristic method (Inact_7, 60.80% and 4.413 metres), the advantage of using a predictive modelling approach is evident.

In terms of classifiers, the neural network model is generally best, closely followed by the adaboost models and random forest. The frequently used logistic regression model scores much lower and actually performs in the same range as the (best) heuristics. The assumptions underlying the logistic model do not seem to fully accommodate the case of home prediction. However, for the models that only use social variables, logistic regression is the second best model in terms of accuracy and even the best in terms of distance error. Taking into account these results and the high interpretability of a logistic regression model, when compared to the other classifiers, this model is advised in a context where only social network data is used.

It is clear that taking all data into account (*full* models) leads to the highest accuracy and the lowest average distance error. The best full model (*full_neuralnet*) performs significantly better than the best model without social network variables (*withoutsocial_neuralnet*) (F-value 33.76, p-value 5.79e-4). The same holds for every classifier in the full model compared with its *withoutsocial* model counterpart. The increase in home detection performance is a bit more outspoken for the weaker logistic regression classifier. In a case where a clearly interpretable model, such as a logistic regression model, is required it therefore becomes even more important to add social variables in order to compensate the loss in performance due to a weaker classifier.

The *socialonly* models obviously can not reach the same performance as all other models that do use the data of the individual itself. Nevertheless, our results do confirm the value of social network data in the context of home detection. In the literature review, evidence was found for home detection in online social networks based on the location of contacts in the social network (Backstrom et al., 2010) and it was anticipated that the value could be much larger in a CDR data set as this is an even better representation of the actual social network of people. Recall that Backstrom et al. (2010) found that the home location was predicted within 40 kilometres of the actual home for 70% of the users. Our best *socialonly* model in terms of distance, the logistic regression based model, lowers this distance at 70% to only 5.2 kilometres. This model is able to classify 95.89% of the data set within the error range of 40 kilometres. The improvement of using CDR data instead of OSN data is substantial. Of course, in order to achieve a fair comparison, one would need to replicate these findings within the same geographical boundaries. Nevertheless, the size of the improvement is a clear indication that a substantial improvement is to be expected.

4.4. Exploratory performance analysis

In order to attain more insight into the performance of the most important models and the performance of home location prediction in general, we develop a further, exploratory performance analysis. We propose seven variables that can be expected to have a relation with the performance measures used in this research.

A first variable *total number of CDR observations* is constructed in order to evaluate the often stated premise that more data leads to better predictions (e.g. Junqué de Fortuny et al. (2013)). Although this largely holds in many situations, one needs to be aware of its possible adverse effect in the case of home prediction. There is a strong, highly significant positive correlation (44.46% p-value < 0.001) between the total number of CDR observations and the number of *distinct towers used*, a second explanatory variable. The consequence of observing more towers for an individual is that selecting the correct tower from the larger set of towers becomes more difficult. This would be very outspoken when using a naive model that randomly selects one of the used towers as the home tower. Although more advanced models will less be affected by this issue, we can still expect a negative relation with performance.

Predicting the correct home tower can also be impeded by the number of towers in the area surrounding the home tower. Urban environments will have a higher tower density than rural areas, which can again make it more difficult to predict the correct tower. The same rationale is the motivation behind the two-step clustering approaches, which are aimed at reducing this problem by clustering nearby towers. Although the results of the clustering methods were unsatisfactory, the rationale still makes sense, which leads to the construction of the following two variables that model tower density. The *number of towers within 2 kilometres of the home tower* can be expected to have a negative effect on performance, whereas the opposite holds for the *distance to the closest tower* to the actual home tower.

The analysis in Section 4.3 indicated the added value of social network data. The *number of contacts* variable is created to assess how it can be more (or less) difficult to predict the home location for people with more social connections.

Finally, we formulate two variables, based on the two best performing heuristic methods, Act_2 and Inact_7. These variables calculate for every individual the total number of counts for these heuristics. For Act_2, this results into *total Act_2 counts*, which is the sum of the distinct days, over all towers used by the individual. This variable captures how many measurements we have to base the well performing Act_2 heuristic on. For Inact_7, this results into *total Inact_7 counts*, which counts the total number of observed inactivity periods, corresponding to the level of inactivity.

	Act_2		Inact_7		Full_neuralnet	
	%Correct Home	Distance Error	%Correct Home	Distance Error	%Correct Home	Distance Error
Total number of CDR observations	-1.30% ***	0.61%	-1.31% ***	0.73% *	-5.58% ***	2.20% ***
Distinct towers used	-5.57% ***	5.76% ***	-5.28% ***	4.82% ***	-10.26% ***	5.93% ***
Number of towers within 2 km of home tower	-11.19% ***	0.33%	-11.43% ***	0.44%	-2.99% ***	-0.79% *
Distance to closest tower	15.89% ***	-1.39% ***	16.30% ***	-1.54% ***	6.77% ***	0.16%
Number of contacts	1.26% ***	-1.25% ***	1.03% **	-1.48% ***	-6.62% ***	0.71% *
Total Act_2 counts	-7.72% ***	2.18% ***	-7.51% ***	1.59% ***	-15.34% ***	3.82% ***
Total Inact_7 counts	2.91% ***	-3.54% ***	3.93% ***	-4.45% ***	-7.53% ***	-0.86% **

Table 8: *Correlations between explanatory factors and performance measures of the best performing methods.* The level of significance is indicated by *** (p-value ≤ 0.001), ** (p-value ≤ 0.01), * (p-value ≤ 0.05).

Table 8 reports the correlations between the seven explanatory variables and the performance of the best methods in the different relevant categories: activity heuristic (Act_2), inactivity heuristic (Inact_7) and predictive model (Full_neuralnet). Note that a positive correlation with % correct home indicates that performance increases when the explanatory variable is higher. However, the opposite is true for the distance error measure, as a higher distance error implies a lower performance. This explains why this correlation usually switches sign, when compared to the corresponding correlation of the % correct home metric.

The results indicate that having more observations for an individual does not lead to higher performance in the case of home detection. The effect that more observations results into more towers clearly explains this finding, as the number of distinct towers has an even

higher negative correlation with the performance metrics. Tower density around the home tower, as measured by the number of towers within a 2 kilometre radius and the distance to the closest tower, has a strong relation with the percentage correct home locations. A higher tower density around the home tower makes it clearly more difficult to detect the correct home tower. The influence of tower density on the distance error is however inconclusive, it is much more limited or even insignificant. This shows that the algorithms still manage to detect a home tower that is relatively close to the actual home tower. As discussed in Section 3.2, this also demonstrates the importance of taking the distance measure into account, as the accuracy alone might underestimate the actual performance in the case of high tower density. A higher number of contacts is related to higher performance for the heuristic methods, however the opposite is true for the predictive neural net model.

A higher number of observations on which the Act_2 (number of distinct days) heuristic can be based, surprisingly leads to lower performance for all models. This result is however perfectly in line with the results of the total number of CDR observations and the number of distinct towers used, all three are measures that to some extent actually model the same concept: the level of activity. In general, the results indicate that it is more difficult to identify the correct home location for more active users. This makes it interesting to have a closer look at the opposite category of inactive methods as well. We observe that the reversed effect holds for the correlation between the *total Inact_7 counts* and the performance of the heuristics. This result is not apparent for the predictive neural net model, where the results are inconclusive. For the heuristic methods, we see that observing more inactive periods, as modelled by *total Inact_7 counts* improves performance. It also has the most pronounced negative effect on distance error for the heuristic methods. In contrast to the level of activity, a higher level of inactivity is positively correlated with a higher performance.

4.5. Combined inactivity activity heuristic method

The benchmark study revealed that the best home prediction method in literature is Inact_5 with 60.69% of the home towers correctly predicted and an average distance error of 4.499 kilometres (see Table 3). An optimisation of the parameter of this approach revealed that these values could be improved to 60.80% and 4.413 kilometres (see Table 5). The main success of the inactivity approach lies in the fact that it effectively models the desired concept of inactivity. This concept was introduced to represent sleeping hours, given that one usually sleeps at his or her home location. Given the data period of five weeks in our research, people would sleep 35 times. In an ideal situation and given perfect data, we would therefore observe 35 inactivity periods for every individual. The total number of inactivity periods is represented by *total Inact_7 count* from the previous section. Figure 4 displays the distribution of this variable. The distribution has the highest relative frequency at 35. The inactivity method therefore clearly achieves to model the desired concept, explaining its strong performance.

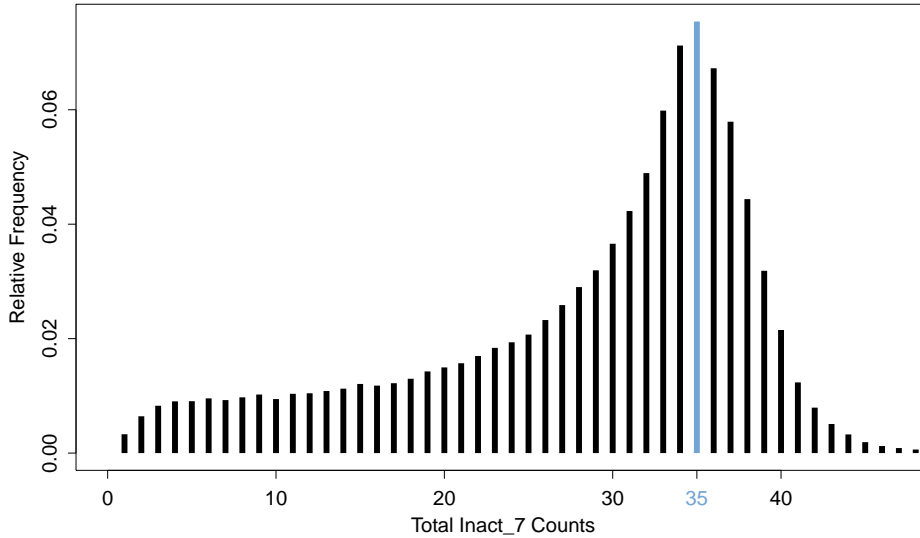
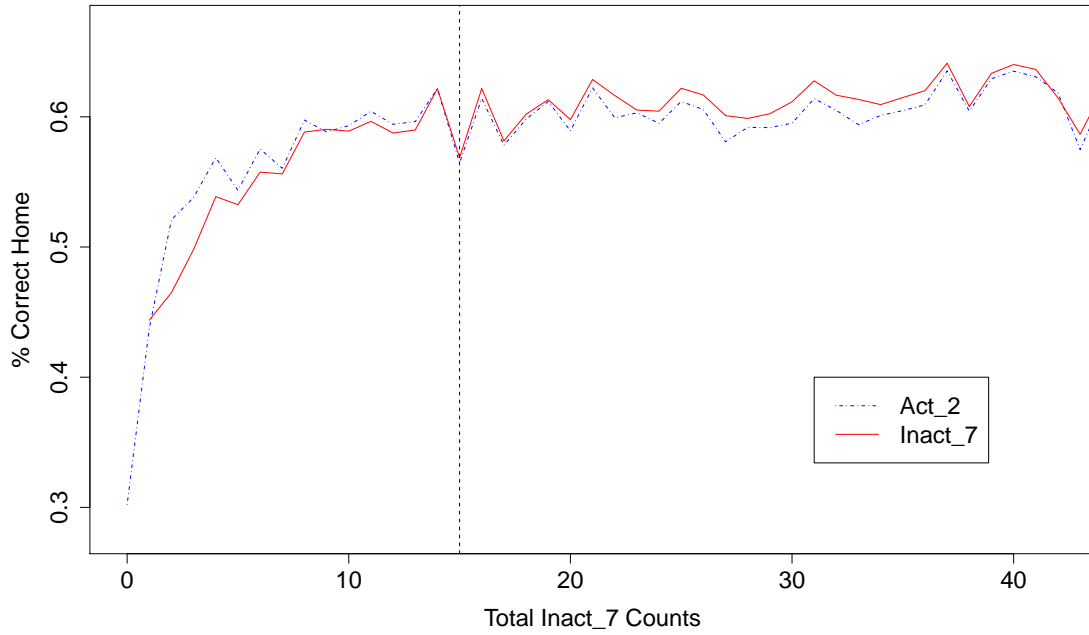
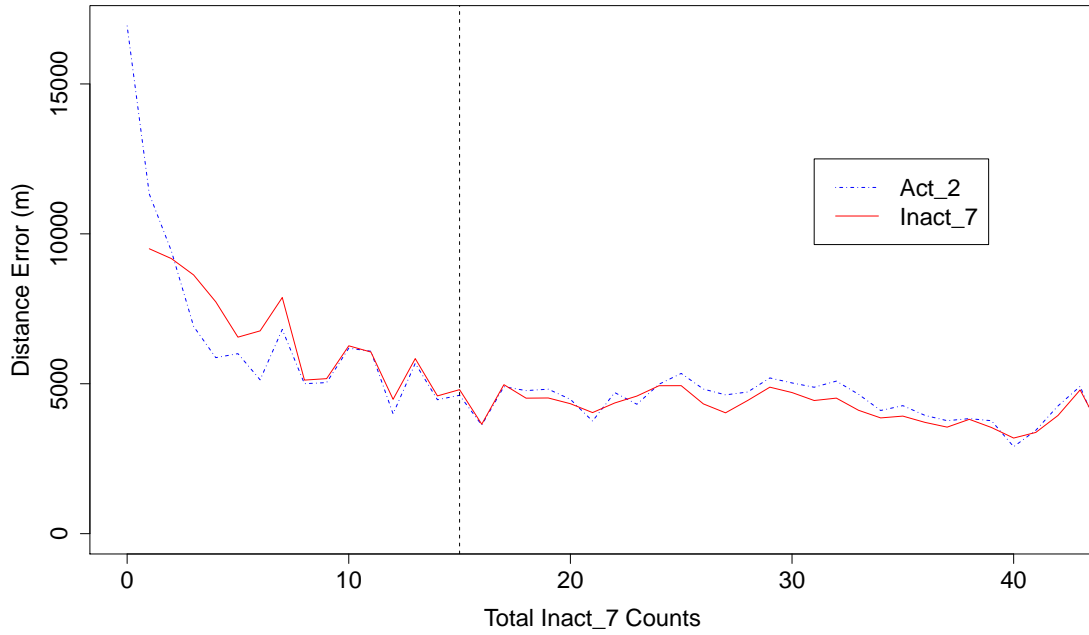


Figure 4: *Distribution of total Inact_7 counts.* This distribution of the total number of inactivity periods is peaked at 35, the number of days in the data set. This indicates that the inactivity heuristic embodies the concept of inactivity, as a proxy for periods of sleep.

Based on the strong performance of Inact_7 and the results in the previous section, a new heuristic has been developed in order to further improve the quality of home detection in an unlabelled setting, where the optimal predictive modelling approach (full_neuralnet) can not be used. The strongest correlation between performance and the explanatory variables was observed for the concept of tower density. Nevertheless, this idea can not be used to further develop a heuristic method for home detection, as the calculation of the number of towers around the home location needs the unknown home location as a prerequisite. The inactivity method seemed to perform best, however, we observed that the performance was positively correlated with a higher level of inactivity. This knowledge hints at the possibility that the inactivity method can be improved when replaced by another, activity based, heuristic at low levels of inactivity. The optimal method to do this is the best activity method, Act_2 (distinct days), which has a slightly lower overall performance (60.16% and 4.591km). The performance of Act_2 has a lower positive correlation with total Inact_7 counts. This combined knowledge indicates that the performance of Act_2 could indeed be higher than Inact_7 at lower levels of inactivity.



(a) Accuracy (% Correct Home) in function of the level of inactivity (Total Inact_7 Counts).



(b) Average distance error in function of the level of inactivity (Total Inact_7 Counts).

Figure 5: *Combined Inactivity Activity Heuristic*. At low level of inactivity, the activity heuristic (Act_2) is advised. At a higher level of inactivity, the inactivity heuristic (Inact_7) is advised. This new heuristic method results in an improved accuracy of 61.00%, and a reduced average distance error of 4.365 kilometres.

These premises are empirically validated and represented in Figure 5. The figures confirm the positive correlations from Table 8 between total Inact_7 and the accuracy (% correct home) for both the activity and inactivity method, as well as the negative correlation with the distance error. Furthermore, the higher correlation for the inactivity method, when compared to the activity method leads to a more rapid increase/decrease, thereby leading to an intersection around the inactivity level of 15. Act_2 does outperform Inact_7 for low levels of inactivity. The figures also indicate that Act_2 always provides a prediction, whereas this is not the case for Inact_7 (0.14% of the individuals have an activity level of zero, leading to no prediction for Inact_7, see also Table 4). We therefore propose to use the Act_2 heuristic for individuals with a low level of inactivity (defined here as total Inact_7 count < 15) and use Inact_7 for higher levels of inactivity (total Inact_7 count \geq 15).

The accuracy rises to 61.00%, while the average distance error decreases to 4.365 kilometers. Applying the 5x2cv F-test informs us that this result is significantly better than both Inact_5, the best heuristic method observed in literature (F-value 157.40, p-value 1.31e-05) and our optimised version Inact_7 (F-value 157.40, p-value 1.31e-05).

4.6. Summary of results

In Figure 6, we plot the reversed cumulative distribution of the distance error for the best performing models for the different categories. Three groups are identified with similar performance. The first group contains the predictive models that only use purely social network variables. It is clear that using merely these variables weakens the results, however these models are interesting if no location data about the individual itself is available and have great potential in such cases. Figure 7 zooms in on the region where the methods start to differ. We observe that *socialonly_neuralnet* has a higher accuracy than *socialonly_logreg*, but nevertheless a higher average distance error, as the neural network model has a larger percentage of users with zero distance error, but is outperformed in terms of distance error quickly as the percentage of users taken into account increases.

A second group consists of the heuristic benchmark methods. Also in that group, we observe that *clust_ha_1_inact5* has a much higher percentage correct home than *clust_ha_5_inact5*, but nevertheless a higher average distance error as well, as the method more quickly increases in the plot.

The third and best performing group consists of the labelled predictive models that use all (*full_neuralnet*) or all, except social variables (*withoutsocial_neuralnet*).

In summarising Table 9, we report the performance metrics (*percent correct home* and *average distance error*) for the best method in each category. Remark that for two of three original benchmark categories, the optimal method is the result of an adaptation in Section 4.2, namely Inact_5 (Dash et al., 2014) was optimized to Inact_7 and the optimal two-step clustering method was Clust_ha_1_inact_5. All original benchmark heuristics are further outperformed by the newly introduced combined activity inactivity heuristic.

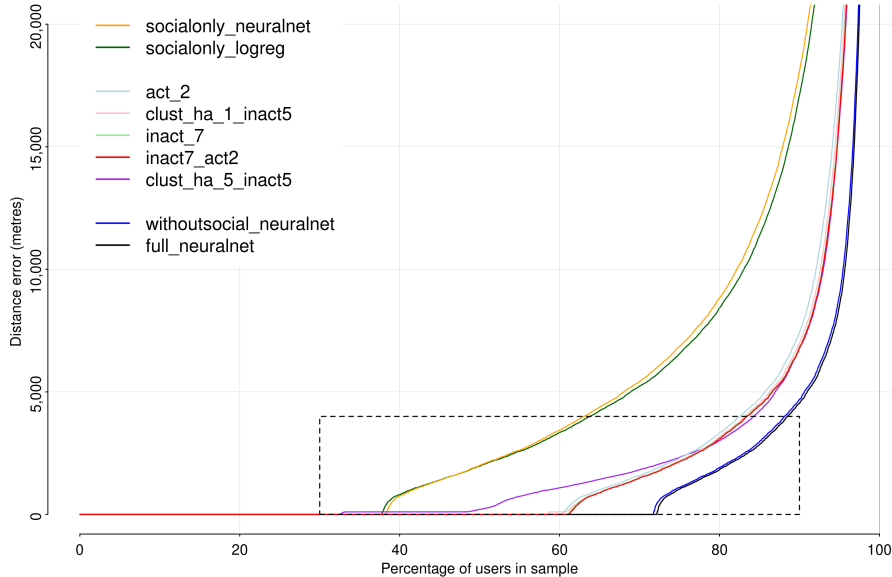


Figure 6: *Reversed cumulative distance error for best performing models in different categories.* The plot displays which error can be achieved for what percentage of the data set. The intersection with the horizontal axis is equal to the *percent correct home* measure. The more a certain method is situated to the bottom right of the figure, the better the performance, as this means that a higher percentage of the data has a lower distance error. Three major groups can be identified in terms of results; the social only models, the heuristic methods and the predictive models. The legend displays the methods from left to right at the upper horizontal line of the dashed rectangle. Figure 7 zooms in on the selected region, bounded by the dashed rectangle, to provide more detailed insights.

Model		Correct home tower (%)	Distance to home tower (km)
Single-step heuristic methods			
Activity heuristics	act_2: Distinct days	60.16	4.591
Inactivity heuristics	inact_7: Inactive periods (7h)	60.80	4.413
Combined heuristic	act_2&inact_7	61.00	4.365
Two-step heuristic methods			
Two-step clustering	clust_ha_1_inact5	58.28	4.525
Predictive modelling methods			
Social only	socialonly_logreg	37.65	8.098
Without social	withoutsocial_neuralnet	71.66	2.941
Full	full_neuralnet	72.08	2.848

Table 9: *Results home detection methods.* The performance metrics (*percent correct home tower* and *average distance error*) are reported for the best method in each category. The Act_2 benchmark is the only benchmark in this figure that was not optimised further, beyond the original method proposed in previous literature. Both Inact_7 and Clust_ha_1_inact5 are adaptations of the original methods. The best heuristic method (act_2&inact_7) is a result of this research. The best overall model (full_neuralnet) improves strongly upon the best heuristic (act_2&inact_7), however is more limited in scope as it requires labelled data.

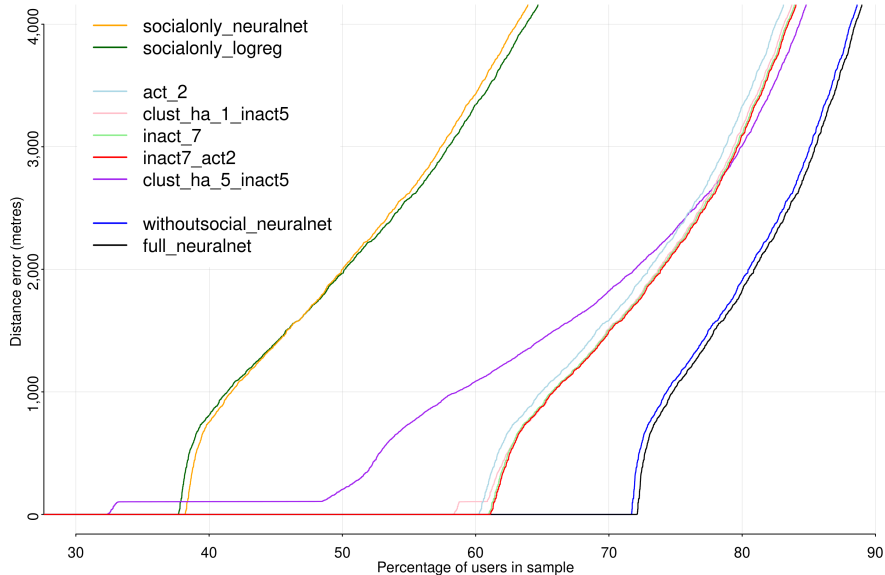


Figure 7: *Zoomed reversed cumulative distance error for best performing models in different categories.* The plot shows a more detailed representation of the performance of the models in the region where they start to differ most. Certain methods combine a lower accuracy (*percent correct home*) with a lower distance error, as they continue below the other method towards the tail of the data set (e.g. *socialonly_neuralnet* versus *socialonly_logreg*).

5. Conclusion and future research

Multiple home detection methods for CDR data have been developed in literature. Nevertheless, there still was a high need for a thorough validation of these methods as ground truth data typically lacks in this type of research. Our unique dataset enabled this benchmark study. The benchmark study revealed that the more complex two-step clustering methods do not lead to higher performance, but on the contrary decrease performance. A second important result is that methods that require the specific choice of a range of hours that define night time perform less good than methods that do not use such parameter settings. The best benchmark heuristics (the distinct days based activity method and the inactivity based method) are both examples of this. The best activity based method relies on the number of distinct days that an individual uses a certain location. This result confirms the research by Vanhoof et al. (2018c). This method does not incorporate the time of the day and is based on the assumption that regularity is an important aspect in identifying the correct home location. The best inactivity method also avoids defining a time of day by modelling periods of inactivity, which aim to model the sleeping hours. These periods can occur throughout the entire day, in order to accommodate for people that work in shift and people with irregular sleeping hours in general. Our analysis identified that the performance of the inactivity approach can be optimised by

changing the required length for the inactive periods. The non-validated arbitrary choice of 5 led to suboptimal results, a longer period of 7 hours improved the performance of this technique. Based on the benchmark results and an analysis of the factors that influence the performance of home detection, we propose a new heuristic that further improves the results in an unlabelled setting. The new combined inactivity activity heuristic uses the activity method for low levels of total inactivity and the inactivity method for sufficiently high levels of inactivity.

Our individual level validation revealed that the best heuristic predict the correct home tower in 61% of the cases and that the average error between the predicted and actual home tower is less than 4.4 kilometres. In an unlabelled setting, where no home locations are known, one needs to rely on these heuristics and the expected performance is as above. In cases where part of the data is labelled, meaning that for a part of the data a home location is known, our research indicates that it is strongly recommended to use a labelled predictive modelling approach. Using a binary classification model enhances the accuracy to more than 72% and reduces the error distance to 2.8 kilometres. The scope of this labelled approach is of course limited to applications with partly labelled data, but this analysis also provides an indication of the maximal attainable performance in home detection when using CDR data.

As a third contribution, we evaluated the value of social network data for home detection in CDR data. We found that adding information about the social network significantly improved the accuracy of the predictive model from 71.66% to 72.08%, and reduced the average distance error from 2,941 metres to 2,848 metres. A predictive model that used merely the data of the social network was able to achieve an accuracy of 37.65% and an average distance error of 8,098 metres, which is a large improvement on previous research using Facebook data (Backstrom et al., 2010).

Identifying the home location is key to many applications, hence our results can be used in a wide variety of research and business applications. Epidemiological models have used CDR data before. CDR data can help modelling the spread of a virus and investigate the impact of measures such as the advice to stay at home during the Covid-19 pandemic. It is obviously crucial to not only have an appropriate data source, but also the accompanying methods to identify what the home location of people in the model is. Previous research (Blondel et al., 2012; Vanhoof et al., 2018b) explained that home detection with CDR data can also be used to replace the outdated or unavailable census data in developing countries. Research about home-work commuting and commuting patterns has a great impact on society by affecting people’s everyday life and by studying the impact of different commute behaviours on our carbon footprint for example. This type of research requires an accurate home location and thus clearly benefits from a solid home detection method. Telecommunications and transportation infrastructure can be further improved and deployed based on findings derived from CDR data analysis. The developed *socialonly* models are of academic interest, but also spark business applications, as these models enable to get insight in the location of non-customers as well. This might lead to identifying areas where the telecom

provider is under-represented and might trigger specific tailored marketing campaigns that boost the customer base of the provider. This way, home detection does not only serve policy makers, human mobility researchers or epidemiologists, but also proves its relevance for marketeers amongst others in business.

We urge researchers to replicate our results on CDR datasets in other countries and cultures in order to assess the robustness in these different settings. Using data of other telecom providers also eliminates a possible selection bias introduced by the fact that a certain provider might attract only a certain segment of the market and therefore produces a biased population sample. This research used five weeks of CDR data, it remains to be investigated how using a longer period of data might affect the results. This aspect is especially important for the proposed new combined inactivity activity heuristic. The concept of using the activity method for a low level of inactivity remains robust. However, the cut-off value where the heuristic switches from the activity to the inactivity method may depend on the length of the observed period. Further research needs to determine whether this value is absolute or relative to the number of observed days. Furthermore, we strongly encourage further research into the promising topic of the models that use social network data, as our preliminary results already reveal great potential. Finally, the ultimate impact of investing in a better home detection method needs to be quantified in the many different applications.

References

- Abualigah, L.M., Khader, A.T., Hanandeh, E.S., 2018a. Hybrid clustering analysis using improved krill herd algorithm. *Applied Intelligence* 48, 4047–4071.
- Abualigah, L.M., Khader, A.T., Hanandeh, E.S., 2018b. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science* 25, 456–466.
- Abualigah, L.M.Q., 2019. Feature selection and enhanced krill herd algorithm for text document clustering. Springer.
- Ahas, R., Silm, S., Järvi, O., Saluveer, E., Tiru, M., 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of urban technology* 17, 3–27.
- Alpaydin, E., 1999. Combined 5×2 cv f test for comparing supervised classification learning algorithms. *Neural computation* 11, 1885–1892.
- Axhausen, K.W., 2005. Social networks and travel: Some hypotheses. *Social dimensions of sustainable transport: transatlantic perspectives* , 90–108.

- Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies* 101, 254–275.
- Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity, in: *Proceedings of the 19th international conference on World wide web*, ACM. pp. 61–70.
- Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. *Physics Reports* 734, 1–74.
- Bharti, K.K., Singh, P.K., 2015. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications* 42, 3105–3114.
- Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. *EPJ data science* 4, 10.
- Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C., 2012. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137* .
- Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., Ratti, C., 2015. Choosing the right home location definition method for the given dataset, in: *International Conference on Social Informatics*, Springer. pp. 194–208.
- Breiman, L., 2001. Random forests. *Machine learning* 45, 5–32.
- Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. *Nature* 439, 462–465.
- Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing* 10, 36–44. doi:10.1109/MPRV.2011.41.
- Carrasco, J.A., Miller, E.J., 2006. Exploring the propensity to perform social activities: a social network approach. *Transportation* 33, 463–480.
- Chen, J., Liu, Y., Zou, M., 2014. From tie strength to function: Home location estimation in social network, in: *2014 IEEE Computers, Communications and IT Applications Conference*, IEEE. pp. 67–71.
- Dash, M., Nguyen, H.L., Hong, C., Yap, G.E., Nguyen, M.N., Li, X., Krishnaswamy, S.P., Decraene, J., Antonatos, S., Wang, Y., et al., 2014. Home and work place prediction for

- urban planning using mobile network data, in: 2014 IEEE 15th International Conference on Mobile Data Management, IEEE. pp. 37–42.
- Dugundji, E.R., Walker, J.L., 2005. Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. *Transportation Research Record* 1921, 70–78.
- Eagle, N., Pentland, A.S., Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 106, 15274–15278.
- Junqué de Fortuny, E., Martens, D., Provost, F., 2013. Predictive modeling with big data: Is bigger really better? *Big Data* 1, 215–226.
- Friedman, J., Hastie, T., Tibshirani, R., et al., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28, 337–407.
- Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature* 453, 779.
- Hartigan, J.A., 1975. *Clustering algorithms* john wiley & sons. Inc., New York, NY .
- Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geolocated twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41, 260–271.
- Hironaka, S., Yoshida, M., Umemura, K., 2016. Analysis of home location estimation with iteration on twitter following relationship, in: 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), IEEE. pp. 1–5.
- Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies* 40, 63–74.
- Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying important places in people’s lives from cellular network data, in: *International Conference on Pervasive Computing*, Springer. pp. 133–151.
- Karikoski, J., Soikkeli, T., 2013. Contextual usage patterns in smartphone communication services. *Personal and ubiquitous computing* 17, 491–502.
- Krings, G., Calabrese, F., Ratti, C., Blondel, V.D., 2009. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* 2009, L07003.

- Kung, K.S., Greco, K., Sobolevsky, S., Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one* 9, e96180.
- Lambiotte, R., Blondel, V.D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P., 2008. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications* 387, 5317–5325.
- Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A., 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences* 102, 11623–11628.
- Liu, F., Janssens, D., Wets, G., Cools, M., 2013. Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications* 40, 3299–3311.
- Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2014. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems* 16, 865–873.
- Mahmud, J., Nichols, J., Drews, C., 2014. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 47.
- Meyners, J., Barrot, C., Becker, J.U., Bodapati, A.V., 2017. Reward-scrounging in customer referral programs. *International Journal of Research in Marketing* 34, 382–398.
- von Mörner, M., 2017. Application of call detail records-chances and obstacles. *Transportation research procedia* 25, 2233–2241.
- Nitzan, I., Libai, B., 2011. Social effects on customer retention. *Journal of Marketing* 75, 24–38.
- Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C., 2012. A tale of many cities: universal patterns in human urban mobility. *PloS one* 7.
- Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L., 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences* 104, 7332–7336.
- Phithakkitnukoon, S., Smoreda, Z., 2016. Influence of social relations on human mobility and sociality: a study of social ties in a cellular network. *Social Network Analysis and Mining* 6, 42.
- Phithakkitnukoon, S., Smoreda, Z., Olivier, P., 2012. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one* 7, e39253.
- Porter, M.F., et al., 1980. An algorithm for suffix stripping. *Program* 14, 130–137.

- R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Roelens, I., Baecke, P., Benoit, D.F., 2016. Identifying influencers in a social network: The value of real referral data. *Decision Support Systems* 91, 25–36.
- Scherrer, L., Tomko, M., Ranacher, P., Weibel, R., 2018. Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science* 7, 19.
- Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010. Limits of predictability in human mobility. *Science* 327, 1018–1021.
- Tang, J., Liu, F., Wang, Y., Wang, H., 2015. Uncovering urban human mobility from large scale taxi gps data. *Physica A: Statistical Mechanics and its Applications* 438, 140–153.
- Tizzoni, M., Bajardi, P., Decuyper, A., King, G.K.K., Schneider, C.M., Blondel, V., Smoreda, Z., González, M.C., Colizza, V., 2014. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology* 10, e1003716.
- Vanhoof, M., Lee, C., Smoreda, Z., 2018a. Performance and sensitivities of home detection from mobile phone data. *arXiv preprint arXiv:1809.09911* .
- Vanhoof, M., Reis, F., Ploetz, T., Smoreda, Z., 2018b. Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics* 34, 935–960.
- Vanhoof, M., Reis, F., Smoreda, Z., Plötz, T., 2018c. Detecting home locations from cdr data: introducing spatial uncertainty to the state-of-the-art. *arXiv preprint arXiv:1808.06398* .
- Vazquez-Prokopec, G.M., Bisanzio, D., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Ramirez-Paredes, J., Halsey, E.S., Kochel, T.J., Scott, T.W., et al., 2013. Using gps technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one* 8.
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L., 2011. Human mobility, social ties, and link prediction, in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, Acm. pp. 1100–1108.
- Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11, 141–155.
- Zagatti, G.A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C.J., Albert, M., Gray, J., Antos, S.E., Burci, P., zu Erbach-Schoenberg, E., 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* 3, 133–165.