Universidad del Desarrollo

Facultad de Ingeniería

Home location detection algorithm comparison using mobile phone data vs real users ground truth

Por: Manuel Antonio Sacasa Ares

Tesis presentada a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al título de Magister en Ciencia de Datos

Profesor Guía: Sr. Leonardo Ferres

Agosto 2020
SANTIAGO

RESEARCH

Home location detection algorithm comparison using mobile phone data vs real users ground truth.

Manuel Sacasa[*]

Correspondence:
sacasamanuel@gmail.com
Master in Data Science, UDD, AV.
La Plaza N°700, Santiago, Chile
Full list of author information is available at the end of the article
[*]Analytical Products Manager.
Telefónica Chile.

Abstract

The detection of Home location is one of the main profiling attributes for telcos, real estate, banking, advertising and targeting companies. Literature shows many examples of criteria, heuristics and algorithms trying to solve the problem. In this document, the objective is to select a final criteria and data source comparing: 5 algorithm, a 2-week data set of three mobile data sources (CDRs-XDRs-Control Plane) and a 75-inhabitants real address data set, built with a survey method, as a ground truth test data set.

This lead us to understand:

- The behavior of the data source: users hit frequency and the impact on algorithm results, method of processing home antenna ranking one, human dependency day-night source behavior and home antenna precision, hit recording logic, noise, error and their limits to solve home location.

- The possible solutions: type of criteria, Geo-time heuristics, counting algorithm types (one or two-step), performance and precision of each algorithm linked to each source.

- The metrics: Test correlation between the closest tower in Euclidean distance versus home tower (output of each algorithm), bias result from 3G data vs. 4g(Lte) antennas, measured Euclidean distance as error between real declared address coordinates and home tower when it's the closest tower or not.

The first step of the document takes the recommendations and implementation of HDAs (Home detection algorithms) and criteria [1][3][5][6][14][16] to compare the difference and behavior between sources and criteria applied to each algorithm (Ranking antenna's hits, ranking antenna's frequency, ranking with time filter, ranking with geographical filter and mix geo-time filters). The second step is to design metrics in order to compare the pair of best performers algorithm with their source: binary metric for a match between HDA result and home antenna ranking, valued match between ranking 1-2-3 home antenna, absolute error distance and MSE distance to real address. The final step is to test an experimental new approach, applying a circadian sleep cycle for each user, to detect time range and process an individual time-range home antenna as a solution for the gaps detected in the first two steps. Comparing circadian results versus the best HDA method, source and groundtruth.

Keywords: home; location; mobile; residence; CDR; XDR; Signalization; mobility; urban

1 Introduction

For Industrial applications, home location is the basic feature for expansion rates, profile enrichment, tokenization with census, profiling tower/antenna Voronoi, improve ML models with geo-social data linked by residence, etc. Home location basically is a token feature to enrich mobility or dynamical data (traces, trips, etc.) whit static geospatial profile descriptive features [6][7].

For academic application, home location is the basic feature to commuter mobility [12][13][14] research, real estate urban mobility, migration between countries or cities, the development of artificial mobile phone census [16] or building metrics for urban planning [8] or politics.

Running experiments and published documents solved the problem with: clusterization [6], counting hits over activity POIs (Point of interest)[7] and counting hits over heuristics and filters assuming POIs, all of these methods over CDRs (Call detail records) [1] or non-continues mobility traces [13]. There is no actual comparison between different cell phone data sources and methods over a real address ground truth dataset.

The proposal is to compare and test (Vanhoof et Al, 2018) HDAs (Home Detection Algorithms) between 3 different mobile phone data sources: CDRs- Call detail records, XDRs data traffic records and Control Plane, over a Groundtruth real address dataset looking for the best performer pair source-algorithm. The best performer pair will be finally tested over an experimental approach call Circadian algorithm solving the real one-on-one overnight sleep time [9][10][15] of each inhabitant looking for home detection.

All the anonymized mobile data belongs from 75 users, inhabitants of Chilean Metropolitan area, specifically from Santiago de Chile, capital city. These 75 users were interview to build a Ground truth data set to test all the results.

2 Related work

Home detection is a basic feature for profiling, urban science, business intelligence and data science research. Is a close feature to mobility patterns with different clues and limitations from detection methods and data sources.

Phone calls data limitations are coming from: human behaviour, data source recording logic and market share. Some limitations are: numbers of calls are almost zero during night time range [13], there's a gender-age difference in the amount and time of calls [17], maximum real geography desaggregation is the tower position or theoretical methods are virtualize antennas positions in grids or Voronoi [1][2][3][13], finally there is a necessary ground-truth data set to validate errors and results [1] were census or independent big surveys cannot be used because of the unknown Telco market share in the studied sample [1][16].

Despite this limitations there is a residential-labor human cycle in labor days and weekends [13] shown in phone data and is possible to create tags for some general topics in mobility traces or dwells ('To work" or "Home")[8]. This tags, home for example, must be related to a POI (Point of interest) that is a place were an inhabitant has daily/weekly frequently activity (Patterns like repeated hits in antennas during certain hours)[14]. One of this patterns is the overnight repeated behaviour describe in [12] as a 24-hours periodicity in human mobility with less locations changed during nights and high predictability. This pattern is shown again in [11] describing 2 stages of profiling during a daily cycle: a low entropy state for home-work location with a few number of antennas and highly predictability and a high entropy state for other activity tags. This low entropy or less information state is describe in [16] as “inactivity” to predict home location. Despite [3][4][8][11][12] describe a 2-kind of land uses (Home-work and other) with different: behaviour, entropy, frequency and predictability, many researchers are focused to developed a single method for tagging all mobility-dwell user’s purpose. This single method [2][3][5][6][7] is based in a phone traces-dwell information and Geo-labeled urban zones measuring activity pattern as hit’s count over cell phone antennas. This two step general method is simplified in [1] as a one step-method using a heuristic approach to labeled ‘home’ activity and focus in the counting methods and accuracy (HDA’s). This heuristic return to the idea of the home-work cycle, low entropy and predictability behaviour call ‘circadian rhythm of the city”[3]. Finally the personal circadian rhythm or sleep pattern is explore in [9][10][15] to understand: if the sleep-awake cycle is altered with age or gender, TST (total sleep-time) behaviour and how to isolate the sleep hour range to process a non-heuristic or real one-on-one overnight pattern to predict home for each inhabitant.

3 Data Sources

Mobile data sources were extracted without pre processing except columns or features basic filters for 75 selected users. The selected users were interview for real address and groundtruth info. The selected sources were chosen by the following criteria:

• Phone call records at antenna level, because it is the mainstream source on literature and urban science [1][12][13][16][17].

• Data transmission records at antenna level . Because it’s a human dependent record (As phone call hits) but with a bigger volume and frequency record. Both sources are stamped or generated by human behavior or mobile usage.

• Machine-to-machine cellphone antenna record. Signalization source with no human dependence or mobile usage between every phone connected to the Telco net and the closest antenna with capacity.

The sources selected were:

1. CDRs (Call detail records) Dataset. Extraction of 2 week history (September 22nd - October 6th of 2019) from 75 users. The features extracted are: phone call in, phone call out, call duration, date time, cell name, district and region.

2. XDRs (Data transmission records) Dataset. Extraction of 2-week history (September 22nd - October 6th of 2019) from 75 users. The features extracted are: phone, amount of Kb transferred, date time, cell name, district and region.

3. Control Plane Dataset (Signalization machine-to-machine). Extraction of 2-week history (September 22nd - October 6th of 2019) from 75 users. The features extracted are: phone, amount of KB transferred, date, time, cell name, district and region.

4. Telecommunication network antenna map. Dataset with cell names, latitude from the towers, longitude from the towers, districts and regions of all antennas in the Telco network. No tilt or azimuth was used.

5. Ground truth dataset (Survey). Phone number and real address from 75-user consent-raised in a human-to-human interview (Fig.5). Only two features were raised in the survey: phone number and exact address (Street name, number and

Figure 1 CDR and XDR. Total number of hits

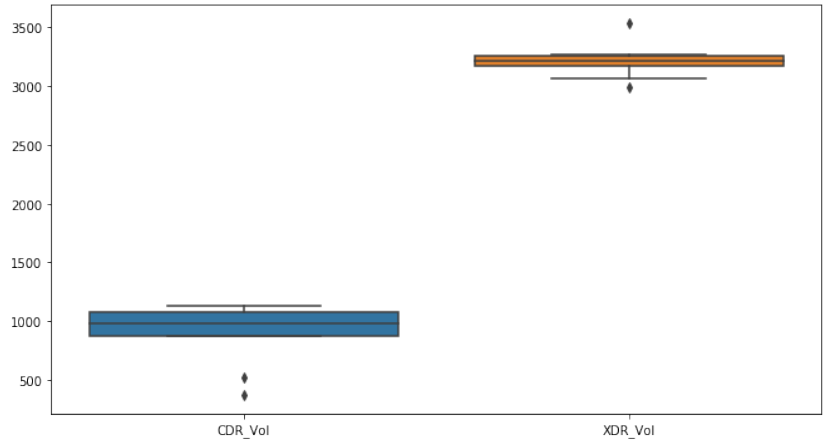
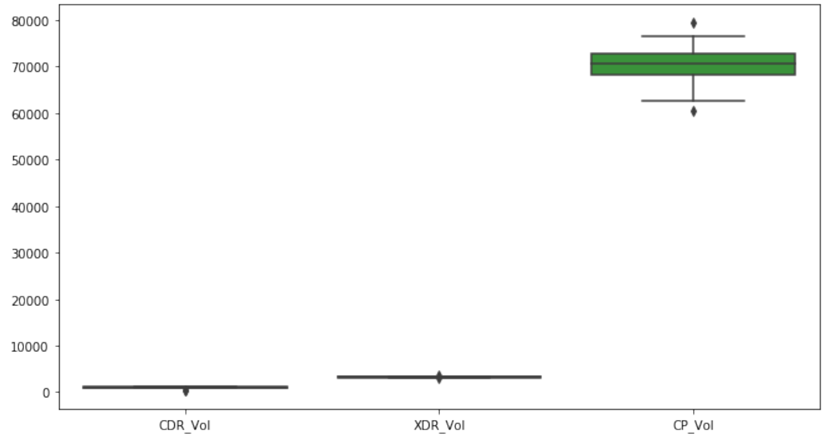


Figure 2 CDR, XDR and CP. Total number of hits



district/comuna. No apartment number was asked). Others features as address coordinates were build one on one to set a high quality standard dataset.

Human dependent sources are distributed as 0.2 CDR hits and 0.8 XDRs hits (Fig.1). For each CDR hit there are 5 XDR hits. The difference between human dependent data and M2M data signalization is 0.05 CDR+XDR against 0.95 Control Plane (Fig.1,2)

Although this sources are human and not-human dependent (CDR-XDR are recorded because of human interaction with their mobile phones) all data sources plots have circadian cycles of daily activity [3]. Only CDRs have a night cycle with 0 hits showing how every hit is related with human activity. XDR, on the other hand, has a sample of hits from automatic activity (Mail update, push messages, app activity, etc.) that holds data > 0 (Fig.3).

Comparing the number of IDs or users in each dataset shows the real nature of each source. CDR has a marked circadian distribution with 0 users during night periods, XDRs a soft circadian [3] distribution with all day baseline > 0 users.

Figure 3 Number of hits per hour

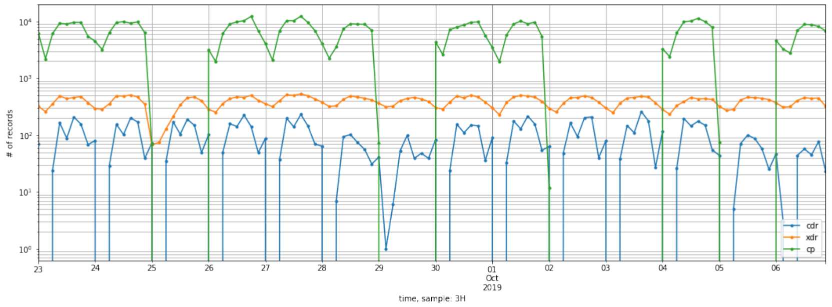
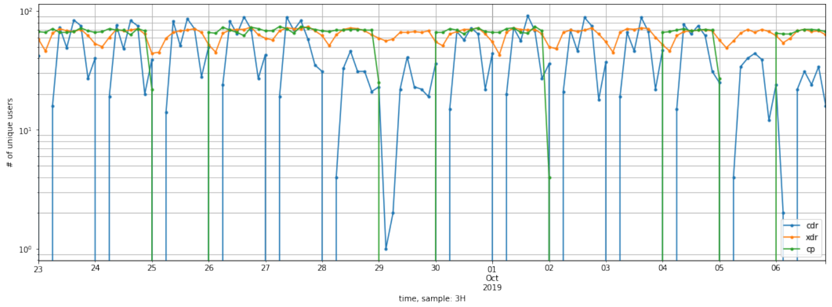


Figure 4 Number of connected users per hour



Finally, Control Plane signalization M2M source records a constant amount of user (Fig.4).

4 Methodology

The methodology was design to run the 5-algorithm with each data source and compare the results over a post-processed groundtruth dataset (Testing dataset). Select the best performers algorithm-sources and compare them with a new approach to conclude the best pair algorithm-source for home location detection. The methodological steps are described below:

Extraction Fifteen days of data (September 22nd - October 6th of 2019) from 75-users interviewed in the ground truth survey. The amount of days were extracted from the three sources directly with no pre processing.

Anonimization All personal data was anonimized with a SHA 256 algorithm transforming to a 64 char strings. Others features as IMSI or IMEI were totally deleted.

Building a testing dataset A Testing dataset was processed with the ground truth dataset and the antenna map. A set of 3-closest antennas were processed for each user, measuring the Euclidean distance between antennas coordinates and real address coordinates. This testing dataset is the main piece to compare results from all the algorithms tested in this document. Describe in section 5.

Running HDAs Each of 5-Algorithm was re-coded to run the experiments and run independently with each dataset (CDR, XDR and control plane) for each user. The results were compared with the closest antenna set on the testing dataset through 2 metrics: a binary match matrix if the home antenna result is/ or not one of the 3-closest antennas and a valued match matrix with 0 for no match and 5,3,1 for nearest ranking. Describe in section 6.

Algorithm comparison With the binary and valued matrix were built a set of 4-errors for finally select the best performer pair algorithm-source: an absolute error, relative error, mean square error and home-mean square error (h-MSE) were calculate in Kilometers with distance between real address and results as a real comparable metric. Describe in section 7.

A new experimental approach The new experimental approach is to determine the real TST (Total sleep time) [15], initial time and final time from each user to process their home antenna. This approach is only possible using a two source (XDR and Control Plane) with human recording logic, To understand sleep range, and non-human logic source to get data during sleep time. The experimental circadian approach will be tested again by the 4-error pool. Describe in section 9.

Finally the document will be concluded with future work guidelines and the results for HDA's + source best performer pair or the improve of the new experimental circadian approach tested over the 75-user ground truth/testing data set.

5 Building a testing dataset

The first step in data processing was the pool of nearest antenna for each user. This result was calculated for each user and each data source in order to build proximity set of possible home antennas. This set of antennas will validate:

- The classic assumption of less Euclidean distance between real address and home antenna in literature (Home antenna/tower is the nearest antenna/tower in a Euclidean theoretical plane)[16].
- The correlation between expert human selections of nearest antennas/towers covering real address coordinates and real address position [11][13].

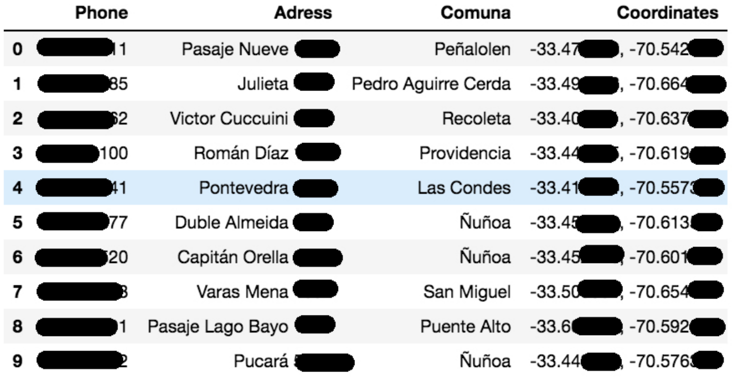
To process the nearest antenna pool, over the entire coordinate list of towers, a K-NN Algorithm with k = 3 (k = number of antennas) was run from the entire tower universe (Fig.6) comparing only Euclidean distance between real address coordinates and tower coordinates (Latitude, longitude) [5].

The final use of a ground truth dataset for Homeloc problem is a quality assurance definition because the blind metrics from high scale testing from census of the unknown market share distribution [1]. The only tool for testing the final results would be a real address of real users. In this case this data is totally necessary to define a solution path for home location.

6 Running HDAs

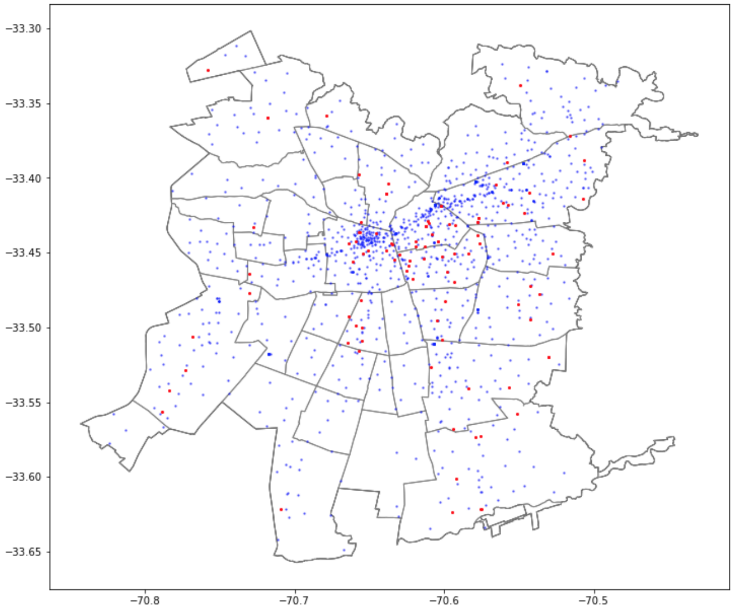
To calculate the home antenna the common decision rules have been used: time limitation criteria based on the assumption of night home time and a regular human behavior [11][12], time based aggregation frequency of hits in the same antenna [6] and spatial geofencing for an area of overnight stay or home behavior [13] resulting

Figure 5 Ground truth survey



	Phone	Adress	Comuna	Coordinates
0	[redacted]1	Pasaje Nueve [redacted]	Peñalolen	-33.47[redacted], -70.542[redacted]
1	[redacted]85	Julieta [redacted]	Pedro Aguirre Cerda	-33.49[redacted], -70.664[redacted]
2	[redacted]62	Victor Cuccuini [redacted]	Recoleta	-33.40[redacted], -70.637[redacted]
3	[redacted]100	Román Díaz [redacted]	Providencia	-33.44[redacted], -70.619[redacted]
4	[redacted]41	Pontevedra [redacted]	Las Condes	-33.41[redacted], -70.557[redacted]
5	[redacted]77	Duble Almeida [redacted]	Ñuñoa	-33.45[redacted], -70.613[redacted]
6	[redacted]20	Capitán Orella [redacted]	Ñuñoa	-33.45[redacted] -70.601[redacted]
7	[redacted]3	Varas Mena [redacted]	San Miguel	-33.50[redacted], -70.654[redacted]
8	[redacted]1	Pasaje Lago Bayo [redacted]	Puente Alto	-33.6[redacted], -70.592[redacted]
9	[redacted]2	Pucará [redacted]	Ñuñoa	-33.44[redacted], -70.576[redacted]

Figure 6 Antenna Map



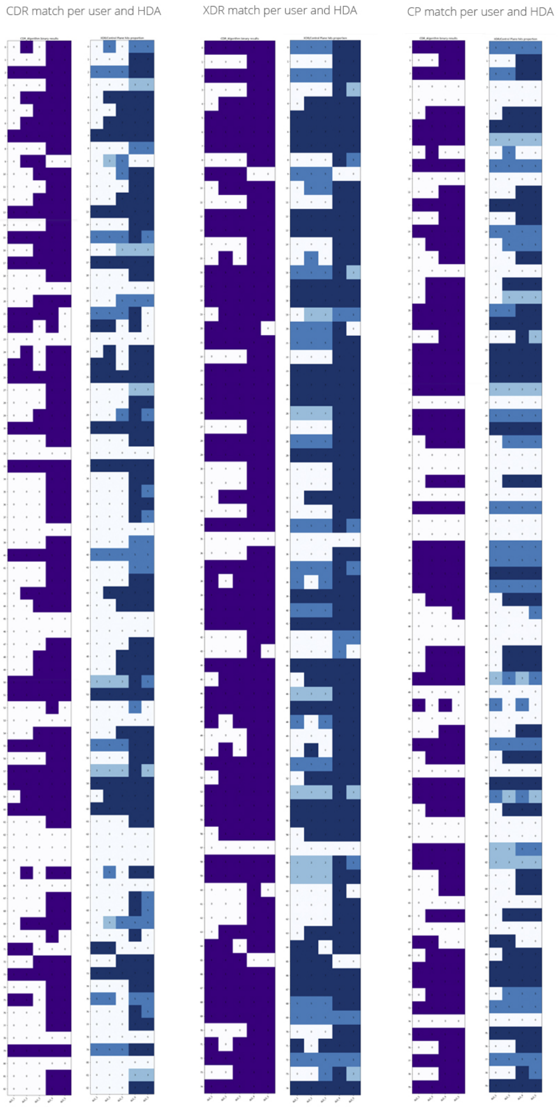
for a multiples interaction with different antennas from the same tower. This decision rules are based over the assumption of a regular human space-time behavior and try to solve the "Single-step approach for non continuous traces" [1]. This approach process a regular place for a space time range but with no real tag of home, that's why a ground truth data set is so necessary [6][7].

To run the single-step experiments, the 5 Vanhoof et al Home detection Algorithms (HDAs) were built. The HDAs are single step approach methods to extract the top 1 ranked antenna with different decision rules. HDAs are:

1. Amount of activity. Ranked antenna list with total Amount of calls in/out for CDR dataset and total amount of hits XDR-CP data set.

2. Amount of distinct days. Ranked antenna list with count of total number of days with total Amount of calls in/out for CDR dataset and total amount of hits XDR-CP data set.

Figure 7 Binary/Valued matrix per source



3. Time of activity. Ranked antenna list with total amount of calls in/out for CDR dataset and total amount of hits XDR-CP data set during a defined time range (Example night or hh:mm-hh:mm)

4. Space of activity. Ranked antenna list with total amount of calls in/out for CDR dataset and total amount of hits XDR-CP data set inside a geofence.

5. Space + Time activity. Ranked antenna list with total amount of call in/out for CDR dataset and total amount of hits XDR-CP data set during a time defined time range (Example night or hh:mm - hh:mm) and inside a geofence.

The HDAs were run over CDR, XDR and CP comparing to k-NN antenna.

Figure 8 Binary match per algorithm

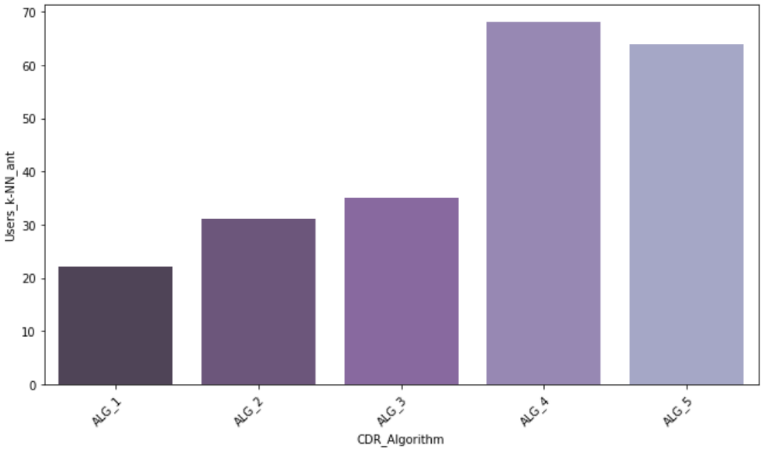
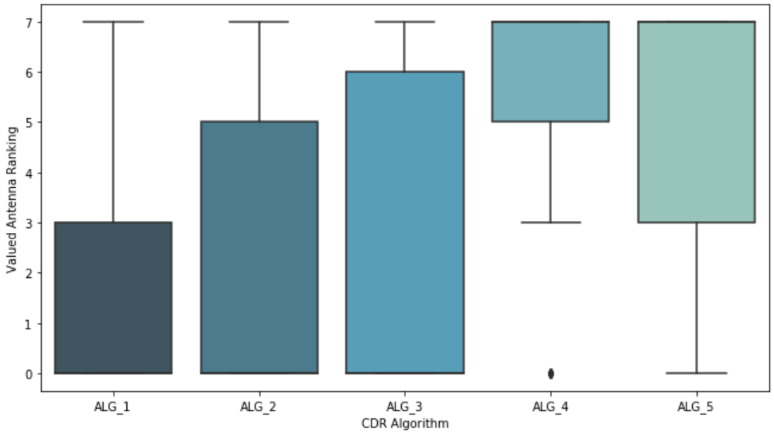


Figure 9 Valued match per algorithm



6.1 CDR Algorithm output

After running HDAs over CDRs, a binary matrix for 1 = match with one of the k-NN antennas, 0 = no match was build. This was the first metric (Fig.7) to visualize the algorithm and source performance.

Table 1 Binary match rate. Home antenna matching one of the three nearest antennas (Fig.7).

	Match Rate
Algorithm N°1 Amount of activity	0.26506024
Algorithm N°2 Distinct days	0.37349398
Algorithm N°3 Time of Activity	0.42168675
Algorithm N°4 Space of Activity	0.81927711
Algorithm N°5 Space+Time	0.77108434

In the binary comparative matrix (Algorithm output match with one of the nearest antenna = 1) "Space of activity" and "Space+time" are the top performers for CDR source. A second comparative matrix was built to measure de difference between type of matches.

For a nearest antenna ranking approach, each algorithm output was compared with the cluster of three k-NN antenna/tower. 3 values were imputed: 0 for no

Figure 10 Binary match per algorithm

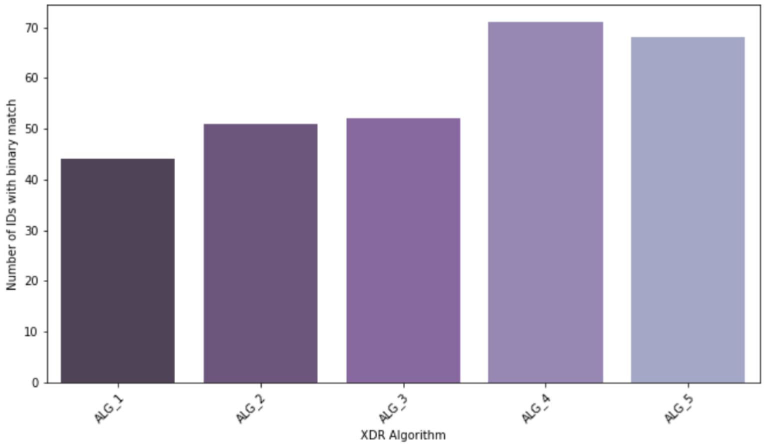
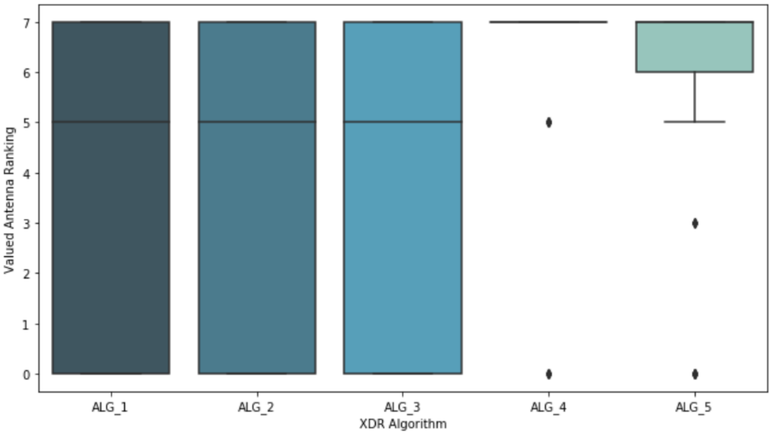


Figure 11 Valued match per algorithm



match, 3 for match with ranking 3, 5 for match with ranking 2 and 7 for match with antenna-tower ranking 1.

Table 2 Valued match rate. Mean of valued nearest antenna/tower type match per algorithm (Fig.9).

	Match Rate
Algorithm N°1 Amount of activity	1.59036145
Algorithm N°2 Distinct days	2.20481928
Algorithm N°3 Time of Activity	2.54216867
Algorithm N°4 Space of Activity	5.3253012
Algorithm N°5 Space+Time	4.77108434

6.2 XDR Algorithm output

The same binary and valued antenna match comprative matrix was built for XDR and CP. The comparison of each Algorithm dataset vs the total amount of users shows the difference between decision rules.

For a nearest antenna ranking approach each algorithm output was compared again with the cluster of three k-NN antenna. 3 values were imputed: 0 for no match, 3 for match with ranking 3, 5 for match with ranking 2 and 7 for match with antenna-tower ranking 1 (Fig.11).

Table 3 Binary match rate. Home antenna matching one of the three nearest antennas (Fig.10).

	Match Rate
Algorithm N°1 Amount of activity	0.58666667
Algorithm N°2 Distinct days	0.6800000
Algorithm N°3 Time of Activity	0.69333333
Algorithm N°4 Space of Activity	0.94666667
Algorithm N°5 Space+Time	0.90666667

Table 4 Valued match rate. Mean of valued nearest antenna/tower type match per algorithm (Fig.10).

	Match Rate
Algorithm N°1 Amount of activity	3.41333333
Algorithm N°2 Distinct days	4.01333333
Algorithm N°3 Time of Activity	4.05333333
Algorithm N°4 Space of Activity	6.49333333
Algorithm N°5 Space+Time	5.92

6.3 Control Plane Algorithm output

Finally, the comparison of each Algorithm dataset vs the total amount of users shows the difference between decision rules.

Table 5 Binary match rate. Home antenna matching one of the three nearest antennas (Fig.12).

	Match Rate
Algorithm N°1 Amount of activity	0.425000
Algorithm N°2 Distinct days	NaN*
Algorithm N°3 Time of Activity	0.6
Algorithm N°4 Space of Activity	0.7125
Algorithm N°5 Space+Time	0.7125

*There were too many days errors in raw datasets to run algorithm 2. It was left apart from the sample.

For a nearest antenna ranking approach, each algorithm output was compared against the cluster of three k-NN antenna. 3 values were imputed: 0 for no match, 3 for match with ranking 3, 5 for match with ranking 2 and 7 for match with antenna ranking 1 (Fig.13).

Table 6 Valued match rate. Mean of valued nearest antenna type match per algorithm (Fig. 12).

	Match Rate
Algorithm N°1 Amount of activity	2.35
Algorithm N°2 Distinct days	NaN*
Algorithm N°3 Time of Activity	3.45
Algorithm N°4 Space of Activity	4.3125
Algorithm N°5 Space+Time	4.2375

Mean of nearest antenna match type value per algorithm. *There were too many days errors in raw datasets to run algorithm 2. It was left apart from the sample.

7 Algorithm comparison

Metrics were designed to compare accuracy of the results. The selection of XDR and CP was made over the preliminary comparison between the binary matrix and the valued match type matrix. In both cases Alg-4 and Alg-5 were the top performers (Fig.8,10,12). RDS (Result dataset) was enriched with Euclidean distance from real address to antenna Top1 (Near0), antenna top2 (Near1) and antenna top 3 (Near2) measured in Kilometers (Fig.14,15,16). This metrics are necessary to build the set of errors to compare and measure the results of each source and Algorithm. The following errors were built for each Id:

Figure 12 Binary match per algorithm

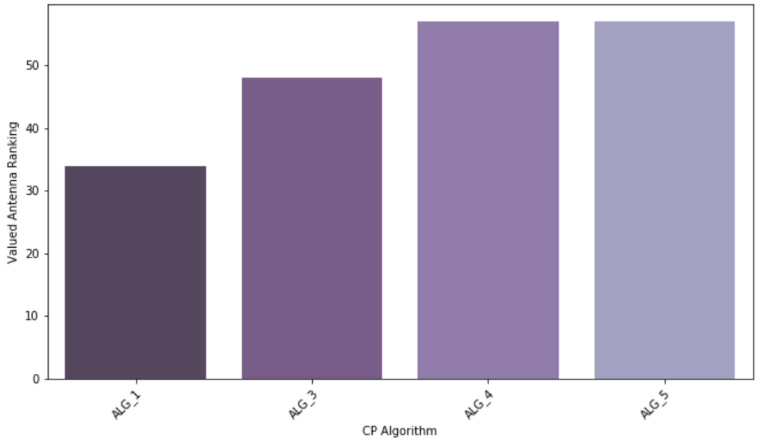
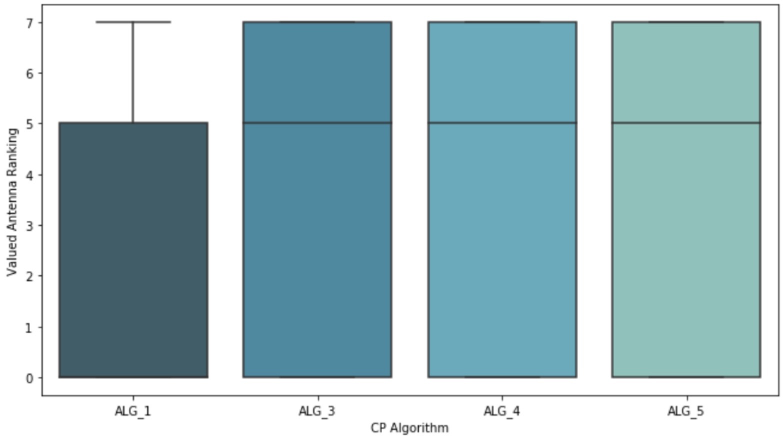


Figure 13 Valued match per algorithm



Absolute error = difference between distance from predicted antenna and real address, and distance from antenna top 1 and real address (Measured in Km).

Relative error = rate of difference between distance from predicted antenna and real address, and distance from antenna top 1 and real address, divided by difference between distance from predicted antenna and real address (Measured in Km).

MSE = Mean square error from the sum of absolute errors

hMSE = Mean square error from the sum of distance between predicted antenna and real address.

The dataset was cleaned from any ID with no algorithm output before running all error metrics. The difference between absolute error means are 0.082590, 0.263737, 0.517545 and 0.567910 Km for XDR algorithm N4, XDR algorithm N5, Control Plane Algorithm N4 and Control Plane Algorithm N5, respectively (Fig.17).

Absolute Error distribution shows how XDR and Control Plane have big differences to capture antenna-ranking 1. Absolute error = 0 or the nearest antenna is a strong metric of performance. The users distribution for Ea = 0 is 67, 56, 35, 34 users (XDR_Alg4, XDR_Alg5, CP_Alg4 and CP_Alg5). Although every Algorithm returns the top 1 nearest antenna, XDR returns two times more amount of Ea = 0.

Figure 14 CDR. Nearest antenna distribution

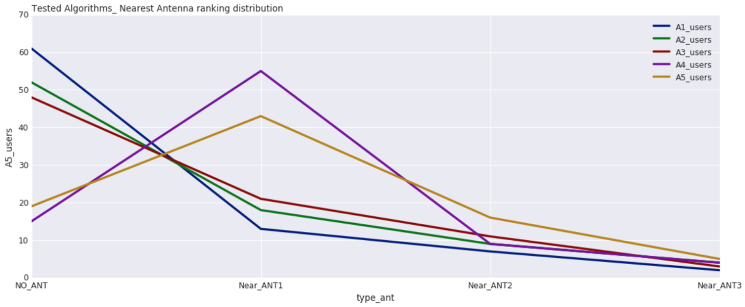


Figure 15 XDR. Nearest antenna distribution

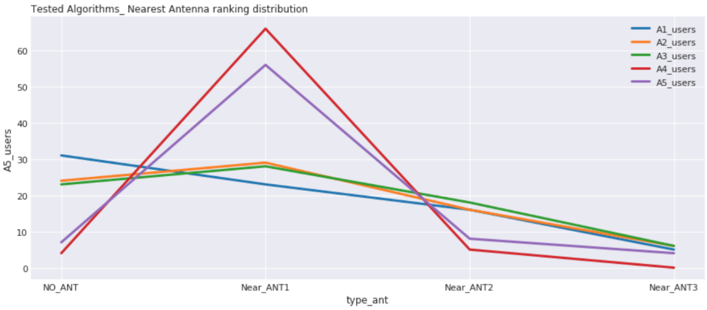
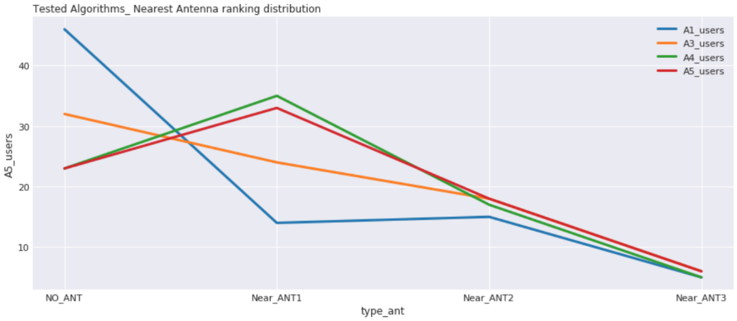


Figure 16 CP. Nearest antenna distribution



For relative error the mean values repeat the difference of absolute error with 0.243614, 0.517595, 1.123972 and 1.270094 km for the list of source/algorithm (Fig.18.

Finally MSE shows more distance between the shown previous results. XDR ALg4 present a small error with almost 0,1% of the higher MSE (Control Plane Algorithm 5). MSE values are 0.025, 0.109, 0.179 and 0.179 for the same consecutive source/Algorithm (Fig.19. hMSE is a weak version from MSE showing less distance between each error. The values 0.239, 0.395, 0.507and 0.516 are closer than regular MSE (Fig.20.

The selected Algorithm-source best performer pair is ALG4-XDR, second place ALG5-XDR, Third place ALG4 and ALG5 with Control Plane (In order).

Figure 17 Absolute error

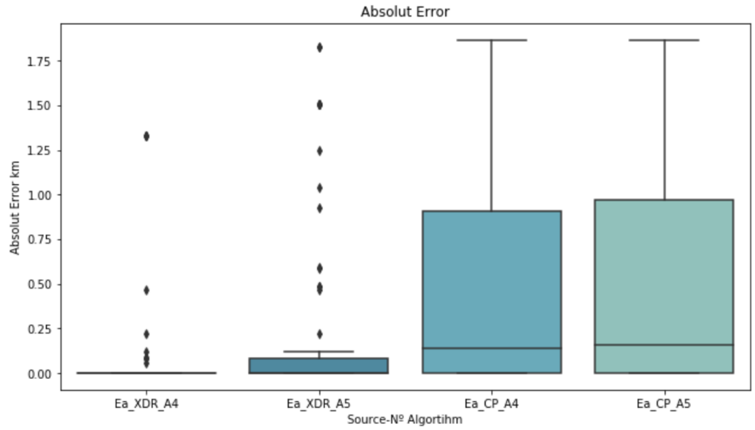
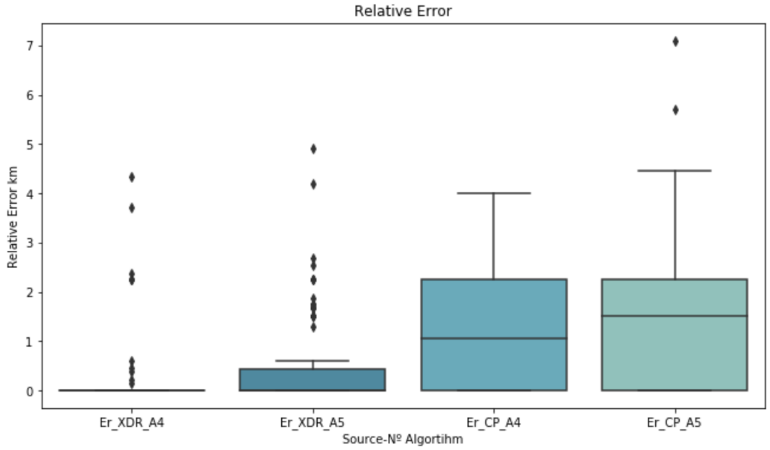


Figure 18 Relative error



8 Limitations and recommendations

The methodology gaps, both for HDA's and multi source comparison with ground truth, are:

• Day type. The dataset must be cleaned or divided between labor days, holidays or weekends. The impact for XDRs may be small hypothetically but the effect over the huge frequency of control Plane is unknown.

• Tilt probability from different sources. The tilt effect is a deterministic jump between antennas following a set of rules from the network caused by a connection collapse or technical issue. Tilt is related to the number of users connected to the antenna and the number of hits. For big amounts of hits there is a bigger probability of tilt. Tilt probability is an unmeasured gap for frequently data source like control plane [16].

• Heuristic time range. Algorithm 4 and 5 are based on a time range for filtering hits over antennas. The values for this time range is an heuristic applied over all users without validation of their behavior. Different home arrival or the number of working hours could mislead the real home [9][10][11][12][15].

Figure 19 Nearest antenna distance MSE

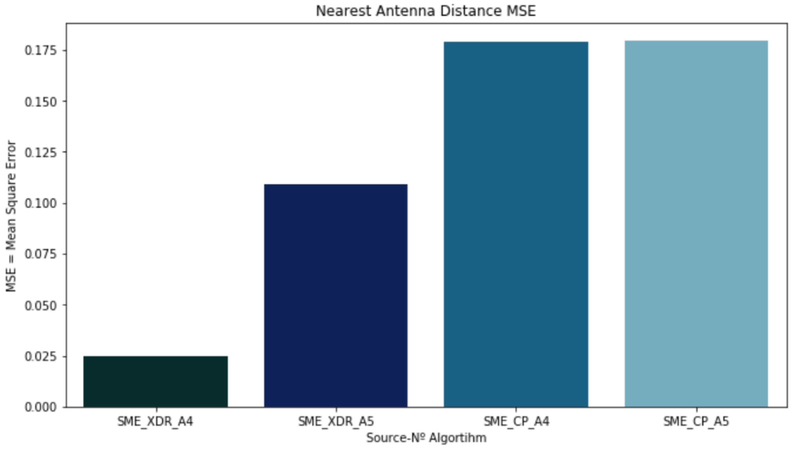
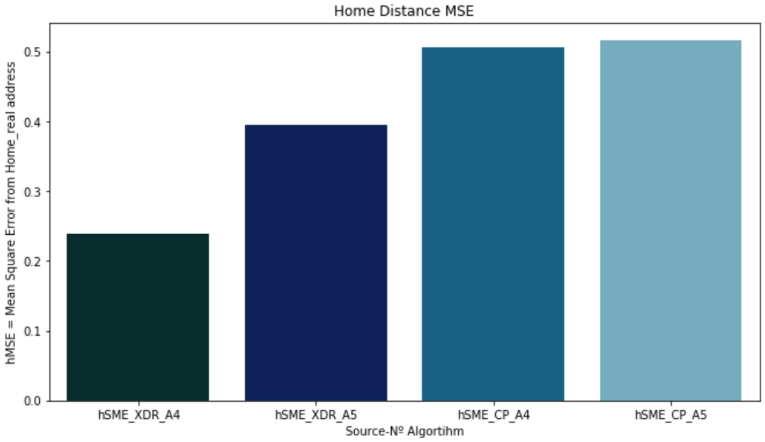


Figure 20 Home distance MSE



• Infrastructure jumps. Calls-in and calls-out trigger an automatic jump from 4g-lte infrastructure (network-antenna) to a 3G network. This network configuration is a big bias for CDRs urban science related to space positions because cell phones will jump to a far away antenna seeking for 3G service [2]. 4G-Lte have more antennas in high density urban zones [11][13].

• Physical urban bias. Linear visual and real coverage area of towers and antennas are deformed by the housing shapes. This might introduce an urban bias in the results of HDAs algorithm. Algorithm could process different datasets from zones clustered by a density index.

From an algorithmic point of view a one-step approach process for all IDs force an automatic heuristic for population (Example: hour range) [12]. Two-step consideration must be made at least to cluster ID's by density [6][11], age [17] or technology consuming profile before running HDA's [1]. ID Clustering IDs could be a simple pre processing solution to minimize possible bias of human behavior .

From the data source point of view, a recommendation is to leave aside CDRs source for home location. It's a strongly circadian human behavior source [13];

Figure 21 Consumption day binary matrix

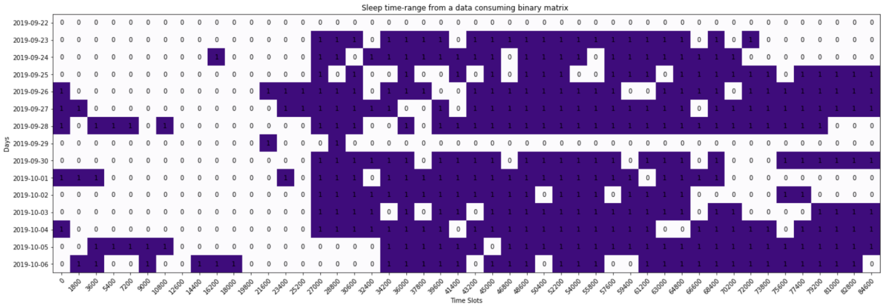
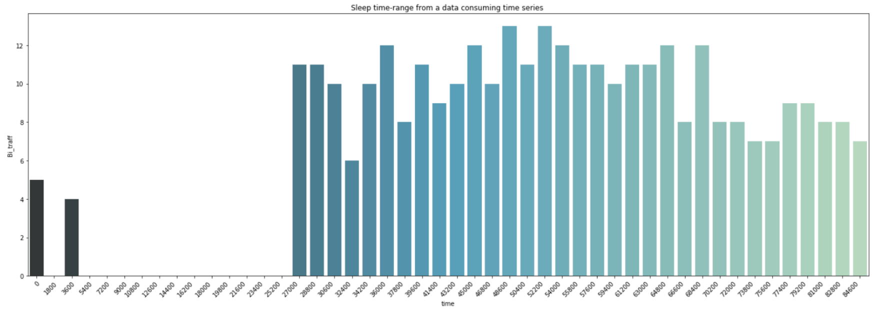


Figure 22 Activity, non-activity pattern



almost non-existing data during night range and infrastructure network 3g-4g jumps make the source inefficient path to arrive to a precise result. CDRs are a perfect source to describe for example the time range for a time detection algorithm but not as de data source to find the home antenna. From this point of view a final idea is to use a combination of sources to strengthen the results.

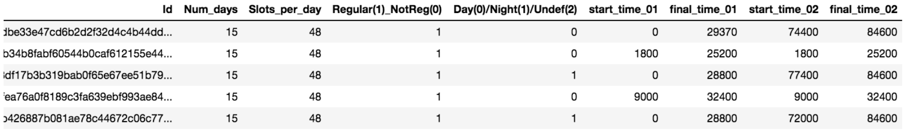
Metrics and errors must switch from a tower latitude-longitude approach to a cell centroid virtual position. This could reduce all distance to real address and reduce MSE.

Finally the ground truth dataset must consider age, gender, urban density index, data consuming profile or index (Heavy, medium, light user) to cluster a two-step HDA approach.

9 Improving Space-time Algorithm

The final experiment is a two-step algorithm testing the possible combination of XDR+CP. The first step is an XDR based algorithm to define the sleep pattern [9][10] for each user raising the real time range to filter data. Home must be defined, as the spaces were an ID overnight so it is imperative to define a one on one sleep pattern. This is only possible using a human behavior [4] data source as XDR. XDR will be the sensor to define the circadian cycle of each user [12][13]. Some users may sleep by day and others may sleep at night, some must be regular, others not, so a nightly behavior must start at 21:00 for some, and others will start at 23:00 [9]. The circadian algorithm raises this differences and distinctions for each user [3]. These are the output features of the CIRCADIAN data set: ‘Id’,

Figure 23 CA result dataset



Id	Num_days	Slots_per_day	Regular(1)_NotReg(0)	Day(0)/Night(1)/Undef(2)	start_time_01	final_time_01	start_time_02	final_time_02
;dbe33e47cd6b2d2f32d4c4b44dd...	15	48	1	0	0	29370	74400	84600
b34b8fabf60544b0caf612155e44...	15	48	1	0	1800	25200	1800	25200
;df17b3b319bab0f65e67ee51b79...	15	48	1	1	0	28800	77400	84600
'ea76a0f8189c3fa639ebf993ae84...	15	48	1	0	9000	32400	9000	32400
ɔ426887b081ae78c44672c06c77...	15	48	1	1	0	28800	72000	84600

Figure 24 Absolute error comparison between HDA's and CA+GCA

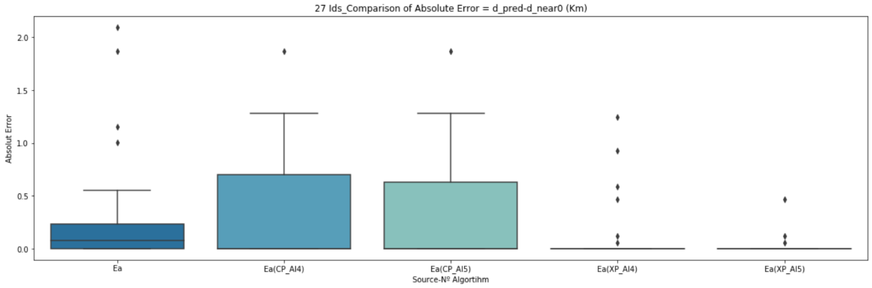
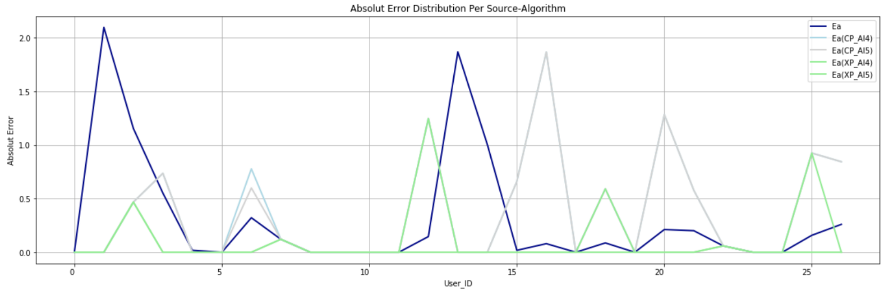


Figure 25 Absolute error per Id



'Num_days', 'Slots_per_day', 'Regular(1)_NotReg(0)', 'Day(0)/Night(1)/Undef(2)', 'start_time_01', 'final_time_01', 'start_time_02', 'final_time_02'.

The CA (Circadian Algorithm) works discretizing time [16] in regular slots and filling them with consuming amount of up/down loaded data (kb). The discrete matrix of time-slot and days is aggregated to build a regular consuming day-type were is possible to detect the real turned off stage of a sleep human with no interaction (Fig. 28, 30, 32, 34).

The CA profiling method returns daily-night sleep behavior describing the non-activity pattern, regularity if sleep pattern has the same frequency during days [4] , TST (Total sleep time) in seconds for the sleep time-range (Fig. 29,31,33,35). This approach avoids a heuristic to describe human behavior and turns to a one-on-one approach [3].

The second-step is named "Geo-gradient" Algorithm (GGA) and is a count method as HDAs but increasing geographic zoom in. Political division of Chilean territory is Region (1rst Scale), Comuna (2nd Scale) and inside Level 2 is the communication network with an antenna or tower division (Voronoi, grid, etc) [6][13]. GGA counts number of hits in each level deleting all the non-top one hits [5]. So It calculates MODE in region level, then deletes all regions except top 1, calculates MODE in comuna level, then deletes non top 1 comuna hits, then selects top one antenna from top one comuna and top 1 Region. This approach avoids tilt between region and comuna. Finally it solves the home antenna with a geo gradient-descent approach [5].

Two experiments were designed to test CA+GGA. The first one combines XDR and CP as data source, XDR as a human sensor and CP because of the volume. The second experiment replaces CP for XDR to test again the performance of sources and criteria. Both experiments CA step were set with the same parameters:

- Minutes = 30. Amount of time for each slot in minutes to measure easily.

- Tolerance = 3. Amount of days with activity considered not regular

- Rand_traff = 500. Amount of Kb considered automatic or not human

- Sleep_hours = 5. Consecutive hours considered sleep-time range.

These are the Circadian steps for both experiments. Circadian Steps:

• Setting a discrete time slot

• Filtering labor-day

• Each day of the analyzed user is divided and filed with XDR data or up/down load traffic.

• An automatic or no human traffic is set (Ex 500 kb).

• A binary matrix with 1=consume and 0 = under the rate .

• The binary matrix is search for the regularity of days slots with value=0.

• All the days are aggregated.

• The array is searched for 5 (Set parameter) consecutive 0s.

• From the array, a nightly time-range or daily time range is filtered in seconds (Ex 00000-27500).

• A dataset with Regularity, day-night pattern, start-end time range is processed.

For GGA there are some definitions described below:

• Morning hour. As the morning final hour for the first period of validate data

• Night Hour. As the night initial hour for the second period of validate data

• List of days. Only labor days from Monday to Thursday.

• Base line. This parameter is hard coded in layers iteration as 0.5 or 50% of hits. It's the selection baseline for weak conditions.

Each layer selection is based on the same strong and weak conditions, described below:

- Strong condition = Layer N top 1 is selected if it gather $N > 50\%$ (baseline) of total N hits (X region is selected if it has $N > 50\%$ of all region hits).

- Weak condition N°1 = If there is no Layer $N > 50\%$, select first pair and if $N1 + N2 > 50\%$ selected top Layer between them (One with bigger amount of hits).

- Weak condition N°2 = If the first pair is $N < 50\%$ but $N1top > N2$ *1.15 (Hard coded parameter) select N1.

- Weak condition N°3 = If the first pair is $N > 50\%$ and $N1top < N2$ *1.15 (Hard coded parameter) select randomly between N1 or N2.

- Neither of the conditions = 0 (No layer)

Finally, a CA result dataset is processed by GGA in a serial model chain. CA sleep hour pattern is taken for each user in GGA input to process home antenna. Both outputs datasets (XDr+CP and XDR) were cleaned from NAN variables and some outliers ending with the following results:

Figure 26 Absolute error comparison between HDA's and CA+GCA (Only XDR)

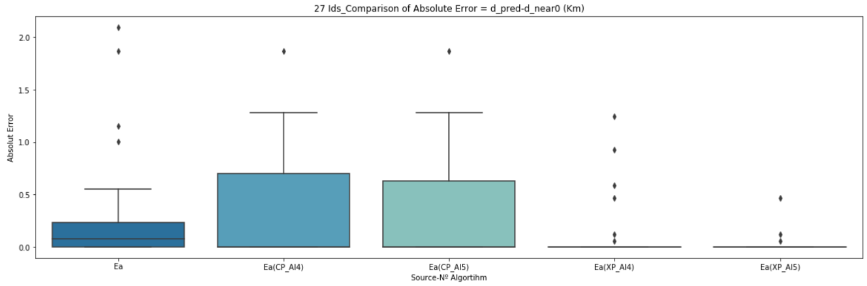
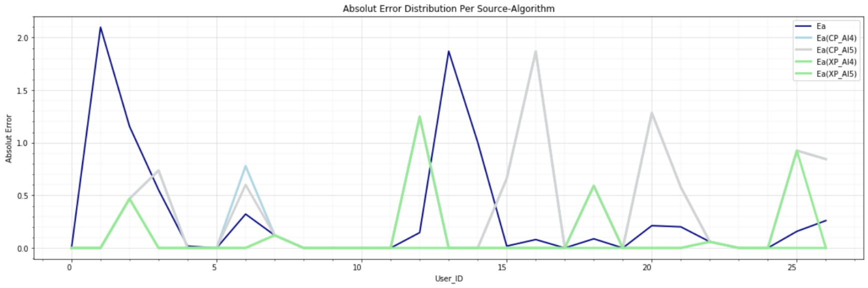


Figure 27 Absolute error per Id. CA+GCA (Only XDR)



None of the experiments perform better than the HDA's 4 and 5 run with XDRs (Comparing absolute errors. Fig.24,26). Both experiments perform better than de CP HDA's (Comparing absolute errors). Experiment 1 (Fig.24,25) run with two sources performs 17,5% better than the exclusive XDR experiment Fig.26,27) comparing mean absolute error (XDR+CP mean Absolute error = 3.09, XDR mean absolute error = 3.62) Standard deviation of absolute error is smaller for exp_1 with x1 =0.564 and x2 = 0.568.

10 Future work

Immediately, after closing experiments and this document, many possible fields of improving, testing and research have been opened. First the signalization/control plane recording and noise behavior. This is a main research for the future of urban science looking to understand the necessary cleaning pre process to use a non bias data source of high frequency and volume. Second, one of the main tests for home location is the comparison between results and census [1]. Of course this is a complex test cleaned from market share and other geographic bias but the final test for a massive location home detection.

11 Conclusions

Despite the fact that home location is a classic problem for non-continuous trace, it seems to be more current than ever. The chance to compare sources and test them against a ground truth dataset has shown some interesting guide lines: Although CDR is the mainstream phone source for research it's the worst source to infer home location. HDAs 4 and 5 (Space and Space-time) perform better with all sources comparing between algorithms. Signalization or control plane, in a raw state in equal pre process conditions than XDR, performs worst despite the difference in volume and frequency. A detailed research of signalization behavior and noise is necessary to understand why XDR is more accurate with 10% of the data. Circadian and Geo gradient algorithm complexity improves CP results but performs worst that HDAs 4 and 5 algorithms with XDR (comparing absolute errors). Heuristics and HDAs may be a better solution combined with XDR than more complex algorithm with Control Plane. Circadian + Geo gradient algorithm returns best results with combined data source than with a single XDR source. This is the only case were XDR performs worse than CP in all experiments presented in this document. This increases the need for a detail research of signalization. All the experiments and explorations must be run with an antenna coverage voronoi scale projecting centroids coordinates. This must return different results from a distance error point of view.

Without a critical analysis of signalization source recording behavior and noise, the conclusion of XDR as the source to detect home location is clear: More complex algorithms do not improve the results compared to HDAs Space and Space-time approach tested against a tower location ground truth.

References

1. Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. Journal of Official Statistics, 34(4), 935-960.
2. Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., & González, M. C. (2013, August). A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing (pp. 2). ACM.
3. Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012, August). Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD international workshop on urban computing (pp. 1-8). ACM.
4. 4. Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. nature, 453(7196), 779.
5. Kang, J. H., Welbourne, W., Stewart, B., & Borriello, G. (2005). Extracting places from traces of locations. ACM SIGMOBILE Mobile Computing and Communications Review, 9(3), 58-68.
6. Yuan, J., Zheng, Y., & Xie, X. (2012, August). Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 186-194).
7. Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011, March). Geographical topic discovery and comparison. In Proceedings of the 20th international conference on World wide web (pp. 247-256).
8. Farrahi, K., & Gatica-Perez, D. (2008, September). Discovering human routines from cell phone data with topic models. In 2008 12th IEEE International Symposium on Wearable Computers (pp. 29-32). IEEE.
9. Duffy, J. F., & Czeisler, C. A. (2002). Age-related change in the relationship between circadian period, circadian phase, and diurnal preference in humans. Neuroscience letters, 318(3), 117-120.
10. Richardson, G. S., Carskadon, M. A., Orav, E. J., & Dement, W. C. (1982). Circadian variations of sleep tendency in elderly and young adult subjects. Sleep: Journal of Sleep Research & Sleep Medicine.
11. Eagle, N., & Pentland, A. S. (2006). Reality mining: sensing complex social systems. Personal and ubiquitous computing, 10(4), 255-268.
12. Song, C., Koren, T., Wang, P., & Barabási, A. L. (2010). Modelling the scaling properties of human mobility. Nature Physics, 6(10), 818-823.
13. Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A. L. (2008). Uncovering individual and collective human dynamics from mobile phone records. Journal of physics A: mathematical and theoretical, 41(22), 224015.
14. Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. Transportation, 42(4), 597-623.
15. Vitiello, M. V. (2006). Sleep in normal aging. Sleep Medicine Clinics, 1(2), 171-176.
16. Dash, M., Nguyen, H. L., Hong, C., Yap, G. E., Nguyen, M. N., Li, X., ... & Anh, D. T. (2014, July). Home and work place prediction for urban planning using mobile network data. In 2014 IEEE 15th International Conference on Mobile Data Management (Vol. 2, pp. 37-42). IEEE.
17. Pearce, E., Launay, J., van Duijn, M., Rotkirch, A., David-Barrett, T., & Dunbar, R. I. (2016). Singing together or apart: The effect of competitive and cooperative singing on social bonding within and between

Universidad del Desarrollo

Facultad de Ingeniería

Home location detection algorithm comparison using mobile phone data vs real users ground truth

Por: Manuel Antonio Sacasa Ares

Tesis presentada a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al título de Magister en Ciencia de Datos

Profesor Guía: Sr. Leonardo Ferres

Agosto 2020
SANTIAGO