8th International Conference on BIG DATA & Data Science for Official Statistics
BILBAO 2024

Informing Climate Change and Sustainable Development Policies with Integrated Data

BILBAO. SPAIN | 10-14 JUNE 2024 | #UNBigData2024

Defining “home” location

Fredrik Eriksson
International Telecommunication Union

“Home” and anchor points

- Determining a subscriber’s “home” is a crucial step for MPD uses cases, e.g. commuting, information society etc.
- “Home” is used to map MPD with reference data, e.g. LAU and population
- A home location in an LAU is assumed to be included in the population estimate for that LAU.
- Other methods include to define more ‘anchor’ points where people regularly stay, e.g. “work”

Many ways to define ‘home’

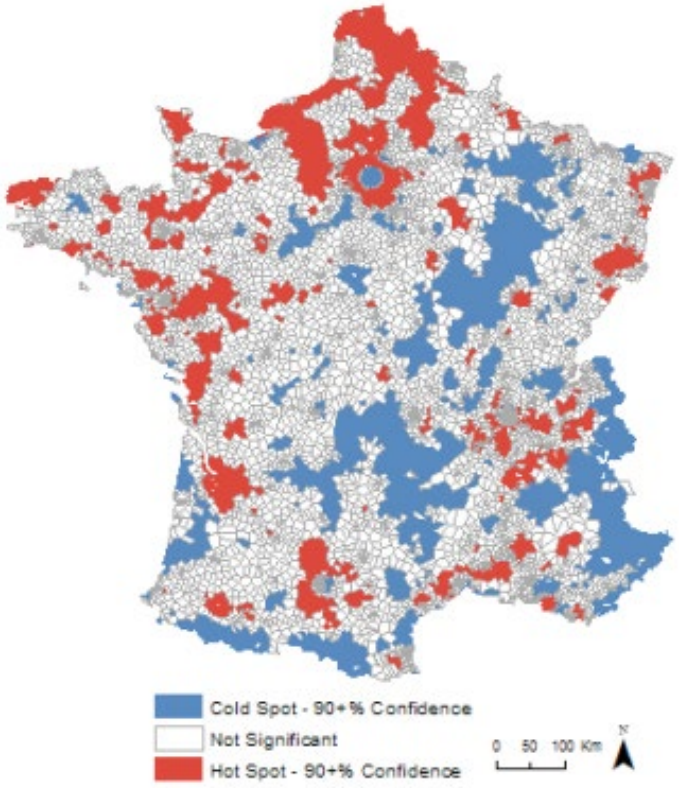
- The amount of activity – “home” is defined as the cell location from where most calls and texts were recorded
- The number of active days – “home” is the cell location from where calls and texts were recorded on the highest number of distinct days
- Time constraints – “home” is the cell location from where most calls and texts were recorded between 7 p.m. and 9 a.m.
- Spatial aggregation – “home” is the cell location from where most calls and texts were recorded within a spatial perimeter, e.g. 1km, around a cell and aggregating all activities within that perimeter
- Combination of time constraints and spatial aggregation.
- More sophisticated models also developed.

Validating ‘home’ algorithms and how criteria used influence the results

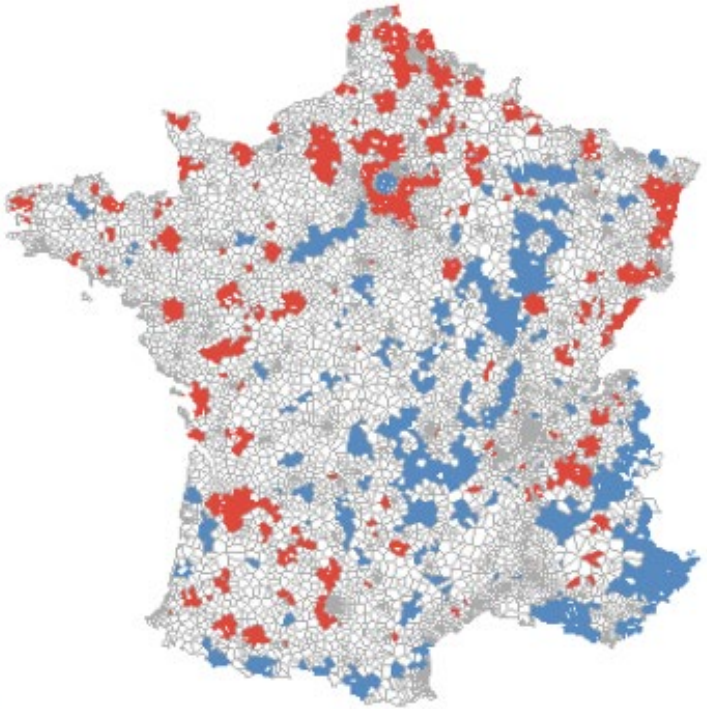
- In Estonia, a “home” anchor point model was up to 99% accurate at the county level and over 90% for higher-level LAU.
- A study of five home algorithms in France showed that the criteria used influenced the detection of home locations for up to about 40% of subscribers.

Avoid summer / vacation season

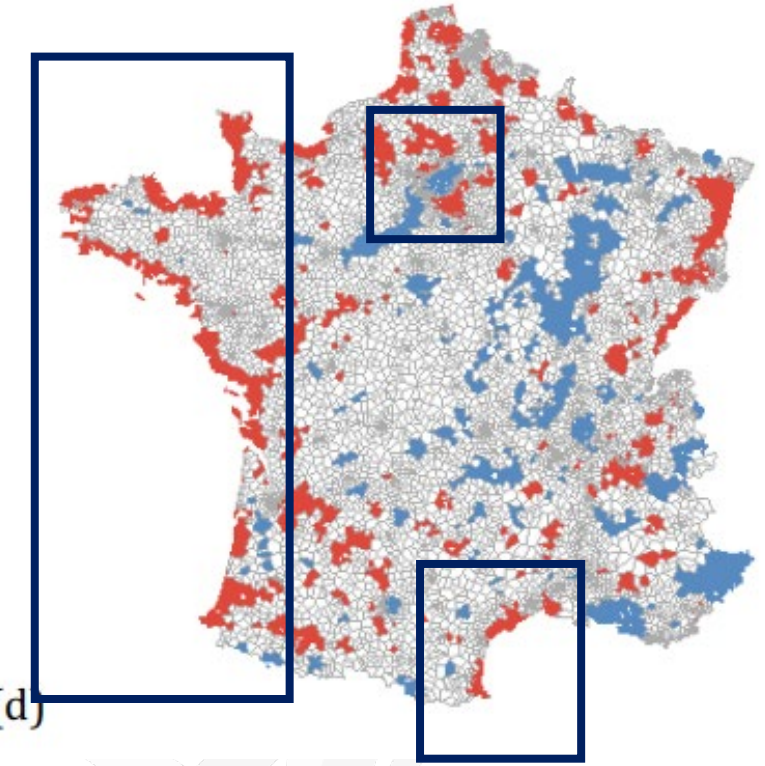
Validation



June



August



Maarten Vanhoof et al., Assessing the quality of home detection from mobile phone data for official statistics

‘Home’ algorithm used in the ITU codes

- Based on Brazil study: a time-constrained “anchoring” model (used to calculate SDG indicator 17.8.1 Proportion of individuals using the Internet)
- The primary objective is to infer the most probable cell location that can be considered the subscriber's home location.
- “Home cell location” identified by:
  - the number events at each cell location from Mondays to Thursdays
  - prioritization within three specific times (Night, Morning, Evening)

User summary - aggregation by subscribers

- The data is summarized into two sets of summary data: User Summary and Cell Summary.
- User Summary calculates user activity statistics, e.g. the number of events/records, unique cell locations, unique days, principal technology used, and highest technology used
- It is needed for indicator calculation to determine whether the user has used the Internet, how often, and by which technology.

msisdn	internet_user	IPDR_events	CDR_events	IPDR_unique_cell	CDR_unique_cell
31	TRUE	635	202	48	15
85	TRUE	2,332	237	76	14
65	TRUE	2,681	73	82	14
53	TRUE	2,910	174	50	10
78	TRUE	1,732	266	61	16

Cell summary - identifying the “home cell”

1. Each event (CDR, IPRD) is classified according to four anchor time categories “Night","Morning","Night“ and "Office Hours".

Default settings:

- Anchor #1 : (00 - 05) -> “Night"
- Anchor #2 : (05 - 08) -> "Morning"
- Anchor #3 : (21 - 00) -> "Night"
- Outtime: (08 - 21) -> "Office Hours"

```
filtered_cell = cell_stats.filter((cell_stats.msisdn == 'subscribers_00007') & (cell_stats.date == '2024-06-10'))
filtered_cell.show()
```

datetime	cell_id	latitude	longitude	data_type	service	date	msisdn	hou	anchor_type
2024-06-10 21:20:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007	2	ANCHOR_3
2024-06-10 19:04:00	2119	43.303	-3.036	IPDR	3G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 19:57:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 05:30:00	2119	43.303	-3.036	CDR	3G	2024-06-10	subscribers_00007		ANCHOR_2
2024-06-10 10:28:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 09:46:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 13:36:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 12:26:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 15:56:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 11:19:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 14:46:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 10:42:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 08:51:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 08:20:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007		OUTTIME
2024-06-10 14:13:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 15:20:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME
2024-06-10 07:15:00	528	43.262	-2.942	IPDR	2G	2024-06-10	subscribers_00007		ANCHOR_2
2024-06-10 13:13:00	528	43.262	-2.942	CDR	2G	2024-06-10	subscribers_00007	1	OUTTIME

Cell summary

2. Aggregate number of events per subscriber, cell and anchor per day

msisdn	date	anchor_type	cell_id	cnt
subscribers_00007	2024-06-10	OUTTIME	528	13
subscribers_00007	2024-06-10	ANCHOR_2	2119	1
subscribers_00007	2024-06-10	OUTTIME	2119	2
subscribers_00007	2024-06-10	ANCHOR_2	528	1
subscribers_00007	2024-06-10	ANCHOR_3	2119	1

3. Assign the most used cell for each anchor and day

msisdn	date	is_weekday	anchor_type	cell_id	cnt
subscribers_00007	2024-06-10	true	ANCHOR_2	2119	1
subscribers_00007	2024-06-10	true	ANCHOR_2	528	1
subscribers_00007	2024-06-10	true	ANCHOR_3	2119	1
subscribers_00007	2024-06-10	true	OUTTIME	528	13

Cell summary

4. For each subscriber – aggregate the number of days for which the cell is the defined ‘anchor time’

msisdn	is_weekday	anchor_type	cell_id	day_cnt
subscribers_00007	true	ANCHOR_1	2119	3
subscribers_00007	true	ANCHOR_2	2119	3
subscribers_00007	true	ANCHOR_2	528	2
subscribers_00007	true	ANCHOR_3	2119	4
subscribers_00007	true	OUTTIME	528	4

5. Infer “home” cell according to multi-step logic

“Direct inference” → **“Tiebreaker”** → **“Indirect inference”**

Determining ‘home cell’ #1/2

1. “Direct inference”

- If the subscriber has one dominant cell during most days during Anchor time 1 (0-5 am), the code assigns the cell to be the subscriber’s “home cell”.
- If the subscriber has two or more cells with the same day count, no “home cell ” is assigned.

Example:

- Out of 30 days, CellID = 123 is the most frequently used cell for Anchor time 1.

CellID = 123 is assigned to be the subscriber’s “home cell”

Determining ‘home cell’ #1/2

2. “Tiebreaker”

- If one of the cells identified during Anchor time 1 is the dominant cell during Anchor time 2 (5-8am), the cell is assigned to be the subscriber’s “home cell”
- If none, the Anchor time 3 (9pm-0am) is checked.
- If multiple other cells are identified to be dominant during Anchor time 2 and Anchor time 3, no “home cell” is assigned.

Example:

- Subscriber “A” has the same number of day count in two cell IDs for Anchor time 1: CellID = 101 and CellID = 936.
- For Anchor time 2, CellID = 936 has more day count than CellID = 101

CellID = 936 is assigned to be subscriber “A”’s “home cell”

Determining ‘home cell’ #2/2

3. “Indirect inference”

- If the subscriber’s “home cell” still can’t be identified by direct or tiebreaker inference, the code randomly picks one location from all the highest frequency candidate locations by prioritizing **Anchor time 1 > Anchor time 2 > Anchor time 3.**

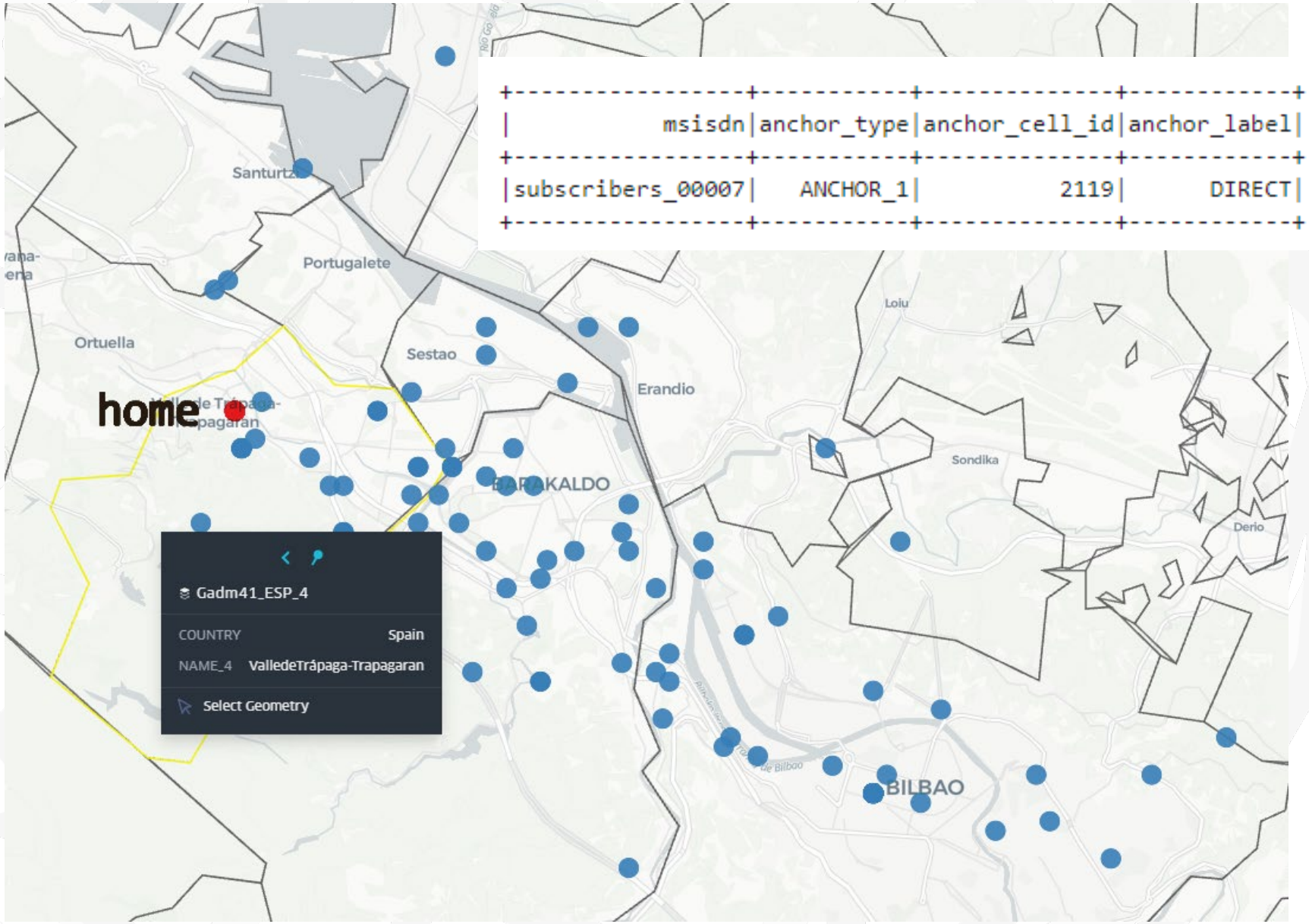
Example:

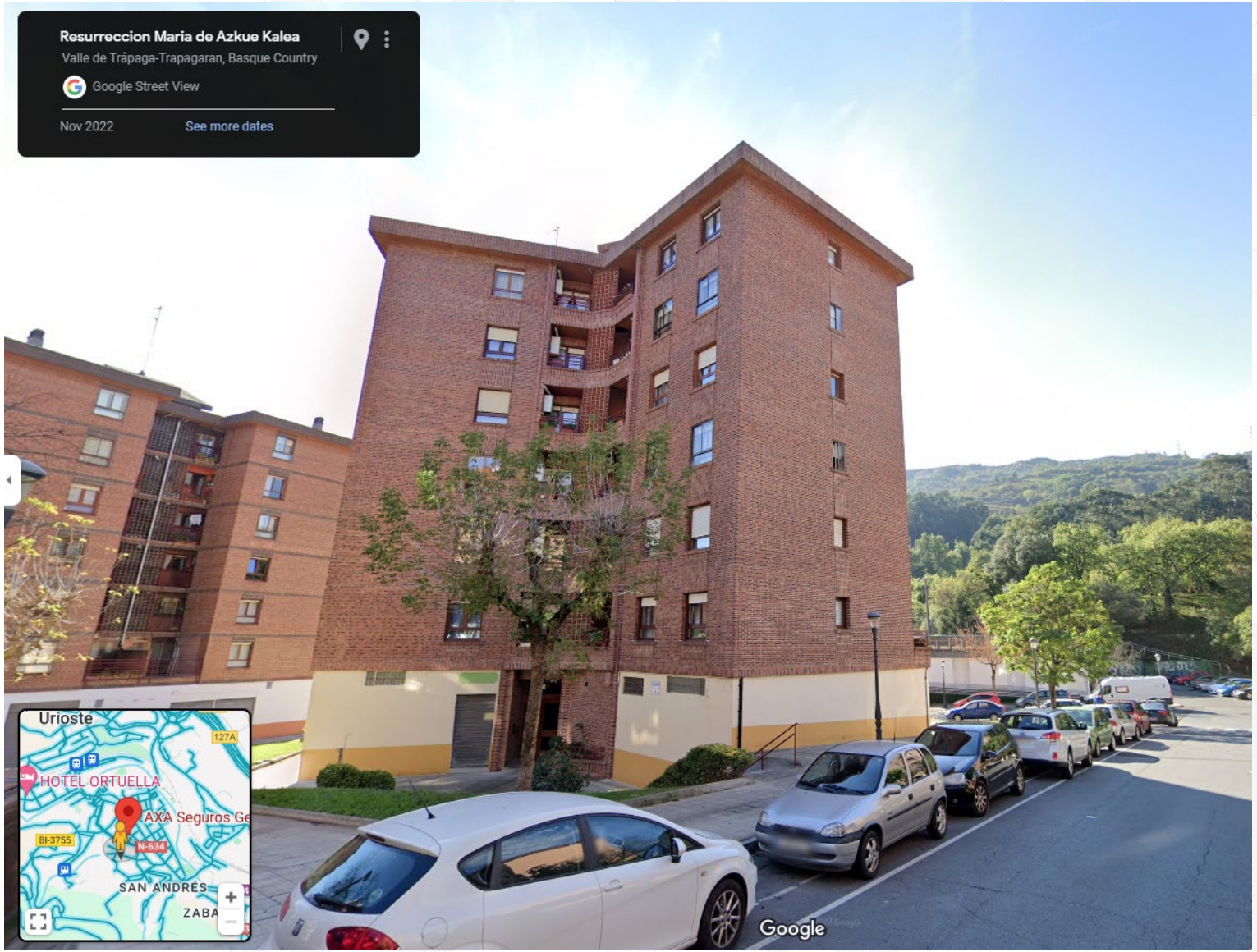
The anchoring result for subscriber “A” is described below:

- Anchor time 1: [{cell_id: “202”, days: 20}, {cell_id: “303”, days: 20}]
- Anchor time 2: [{cell_id: “505”, days: 18}, {cell_id: “404”, days: 18}]
- Anchor time 3: No Data

CellID = 202 from Anchor time 1 is chosen as subscriber “A”’s “home cell”

Example


+----------------+----------+-------------+-----------+
| msisdn|anchor_type|anchor_cell_id|anchor_label|
+----------------+----------+-------------+-----------+
|subscribers_00007| ANCHOR_1| 2119| DIRECT|
+----------------+----------+-------------+-----------+
Santurtzi
Portugalete
Ortuella
Sestao
Erandio
Loiu
Sondika
Derio
home
BARAKALDO
BILBAO
Gadm41_ESP_4
COUNTRY Spain
NAME_4 ValledeTrápaga-Trapagaran
Select Geometry


Resurreccion Maria de Azkue Kalea
Valle de Trápaga-Trapagaran, Basque Country
Google Street View
Nov 2022
See more dates
Urioste
127A
HOTEL ORTUELLA
AXA Seguros Ge
BI-3755
N-634
SAN ANDRÉS
ZABA
Google

Next: Indicator calculation

United Nations BigData

EUSKO JAURLARITZA GOBIERNO VASCO

Bizkaia foru aldundia diputación foral

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS EXTERIORES, UNIÓN EUROPEA Y COOPERACIÓN

BC3 BASQUE CENTRE FOR CLIMATE CHANGE Klima Aldaketa Ikergai

EXCELENCIA MARÍA DE MAEZTU 2023-2027

INE Instituto Nacional de Estadística

Eustat EUSKAL ESTATISTIKA ERAKUNDEA INSTITUTO VASCO DE ESTADÍSTICA

#UNBigData2024