






Call Detail Records for Human Mobility Studies

Taking Stock of the Situation in the “Always Connected Era”

-  Pierdomenico Fiadino (EURECAT, Spain)
-  Victor Ponce (Eurecat, Spain)
-  Juan Antonio Torrero (Orange Spain)
-  Marc Torrent (Eurecat, Spain)
-  **Alessandro D’Alconzo (AIT, Austria)**

Big-DAMA 2017, Los Angeles

What are CDR



CALL DETAIL RECORDS (CDRs)

CDRs are tickets to **support operators' billing procedure**, summarizing a transaction, such as a call, text message or data connection.



Fields:

User IDs (anonymized*)

Timestamp

Cell info (ID, type, **coordinates**)

Action type

Others (routing, bill data, etc.)

Types of recorded transaction:

calls (inbound and outbound)

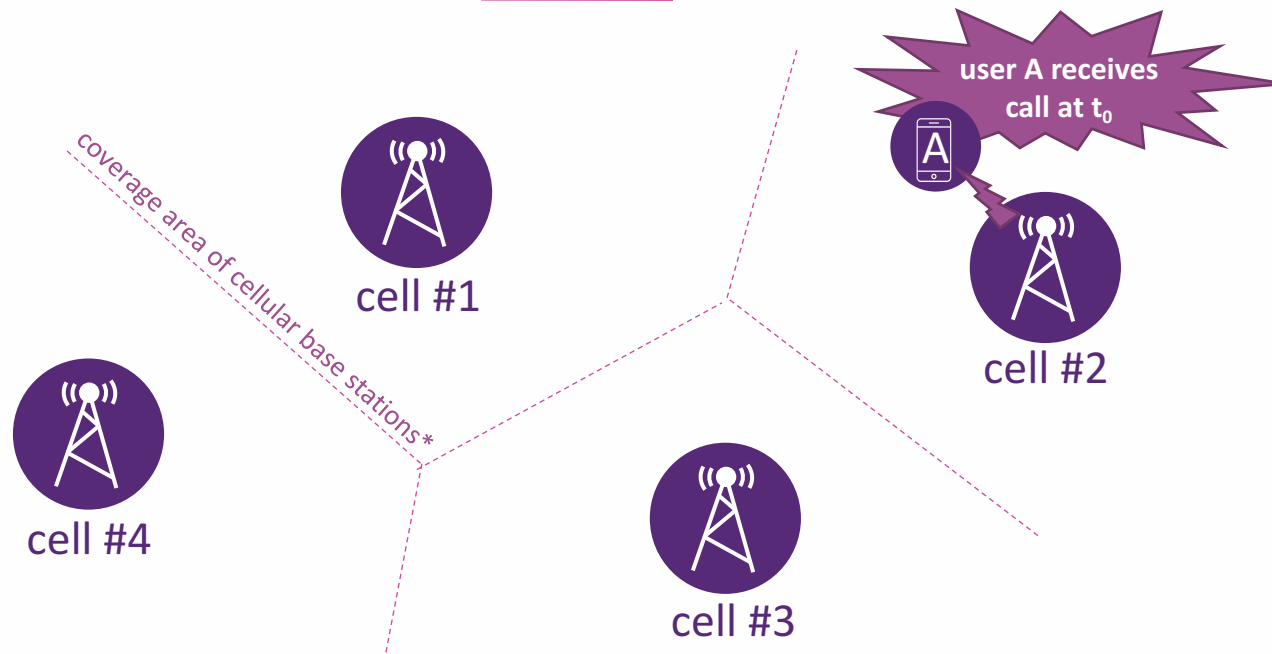
SMS (inbound and outbound)

data connection (beginning of PDP context)

*still, user metadata might be available (age, gender, post code, etc.)

CDRs and Human Mobility

Geo-localization of user actions



example of a transaction

user_id	timestamp	action_type	cell_id	...
A	t_0	inbound_call	2	...

user_id	age	gender	rate	address
A	30	M	MyRate™	Rambla, BCN

dimension (user-related metadata)

cell_id	type	latitude	longitude	...
2	micro	40.9220	1.7612	...

dimension (cell-related metadata)

* The actual radio planning is in general more complex (cell sectors, overlapping areas, umbrella cells, etc.)

CDRs and Human Mobility

Opportunities and Caveats

- The localization of actions allows reconstructing human mobility
- Support large-scale studies of aggregated behaviors
- Wide range of applications (socio-economical studies, transport optimization)
- Rich data (not only positions, but also demo-graphic data: gender, nationality..)
- Rather high data availability (depending on operator's exploitation strategies)
- Rich literature (preliminary studies/proof of concepts, *but no real-world use, yet*)

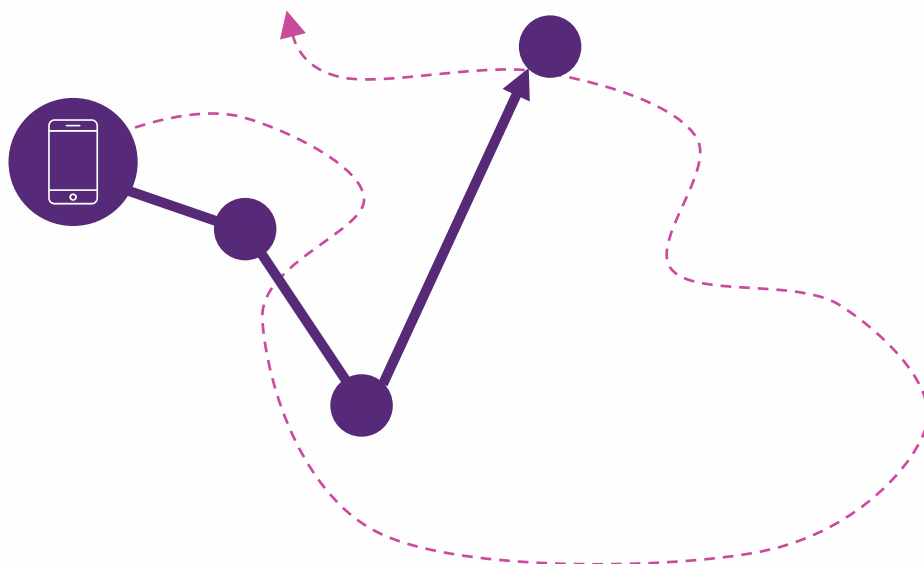
Caveats:

- Designed for different purposes (i.e., billing)
- Variable space granularity (location accuracy, depending on cell towers)
- (Historically) low time granularity (umber, frequency and uniformity of samples)

CDRs and Human Mobility

Time and Space granularity: actual mobility vs. perceived mobility

- ➔ Actual user/handheld trajectory
- Recorded actions (antenna position)
- ➔ Perceived trajectory (from CDRs)



Location accuracy depends on cell position and radio planning

Usually not a big issue, depending on applications (urban radio planning is usually rather *dense*)

Main CDR limitation:
limited time granularity

User's location only "visible" when an action occurs

Low time granularity = low accuracy in tracking users

Old literature shows that CDR are characterized by low time granularity. Is that changed?

Premise: past literature highlights the limited time granularity of CDRs

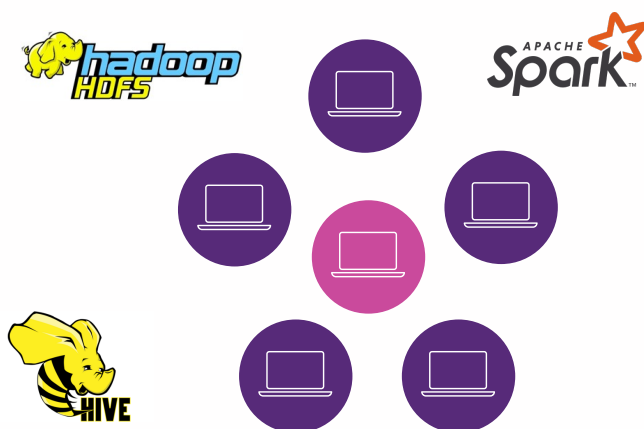
Question: has this changed?

OUR GOAL

Characterize the quality of
mobility information provided by CDRs
considering nowadays cellular usage patterns

Hard-facts and infrastructure

	DS2014	DS2016
Collection period	Q3-2014	Q2-2016
Geographical area	Nation-wide (Spain)	
Length	31 days	
# records/day	350 million	1.1 billion
# users/day	9 million	11 million
Data volume/day	50GB (compressed)	120GB (compressed)



6-nodes cluster (1 master, 5 workers)

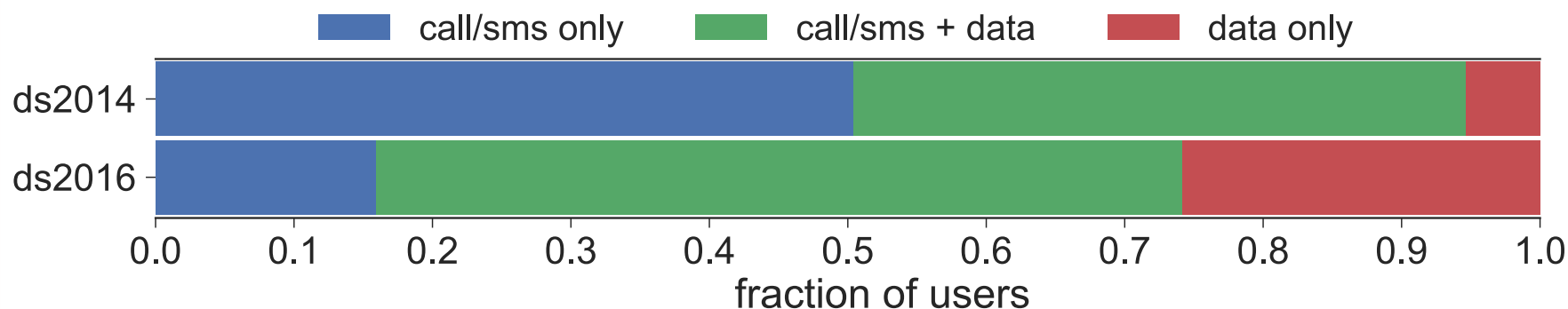
48 CPUs, 192GB memory, 10TB storage

Hadoop Distributed File System (HDFS)

Apache HIVE (for ETL and pre-processing)

Apache Spark (for data analytics)

Cellular traffic type, from 2014 to 2016



Radical change in usage patterns: more data connection and “always on” apps

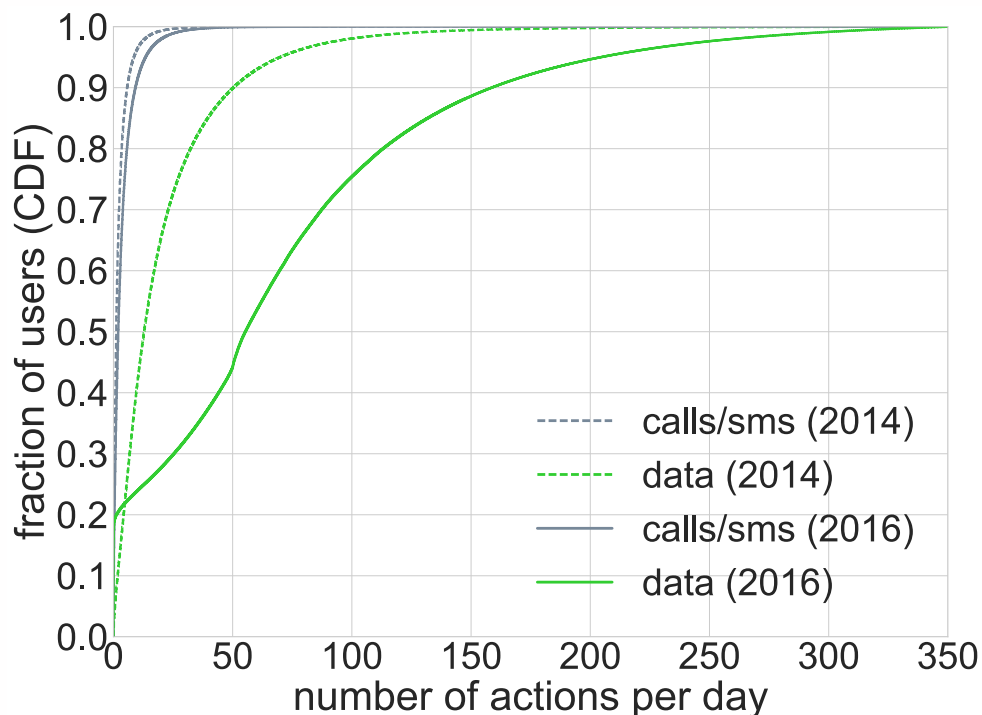
Data connection users: from ~50% in 2014 to ~85% in 2016

2G/3G/4G users generate more actions and more frequently



Increased action rate

# tickets	ds2014	ds2016	Δ
SMS/day	0.5	0.1	-0.4
CALLS/day	1.8	2.5	+0.7
DATA/day	10.9	50.1	+36.9
TOTAL/day	13.2	52.7	+39.5



Key observations:

- SMS** became marginal (replaced by instant messaging apps over 3G/4G)
- Small increase of **voice** calls (increment of flat rates, restrained by VoIP)
- Extremely high increase of **data**-related tickets (even considering data-users only)
- Overall increase of average number of records per day per user: 50% of users generate at least 50 actions per day (only 10% did the same in 2014)**



Data quality indicators



Definitions

DAYS OF ACTIVITY (DOV)

fraction of days in which a user generates actions over entire observation period

the higher the better

HOURLY ACTION RATE (HAR)

average number of action per hour

the higher the better

AVERAGE LAG TIME (ALT)

average gap between two consecutive actions

the lower the better

TOTAL INACTIVE TIME (TIT)

How long a user is idle (no recorded CDRs) over the observation period

the lower the better

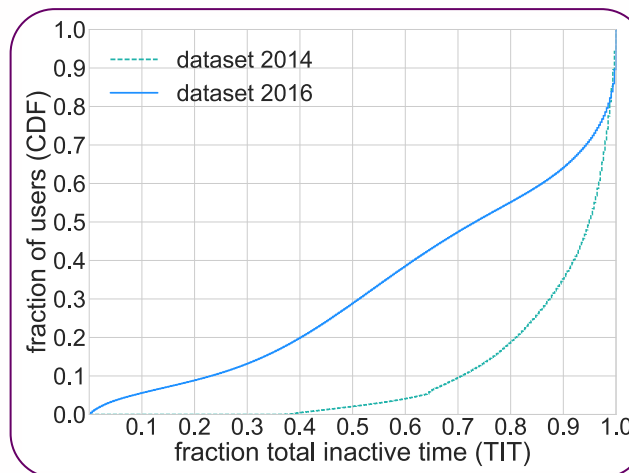
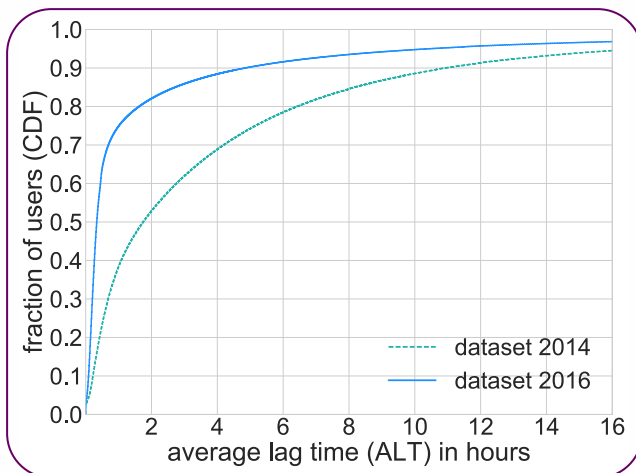
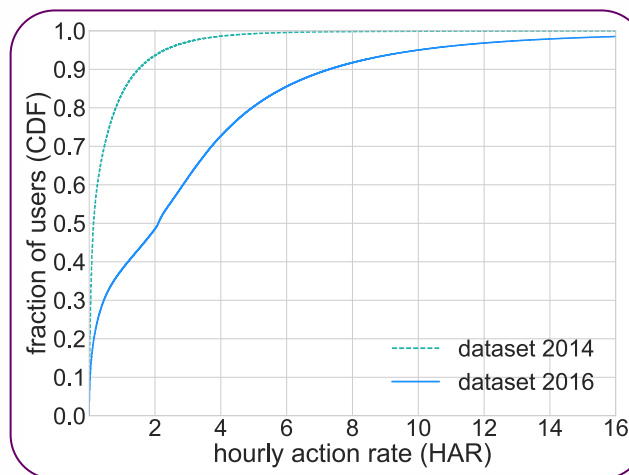
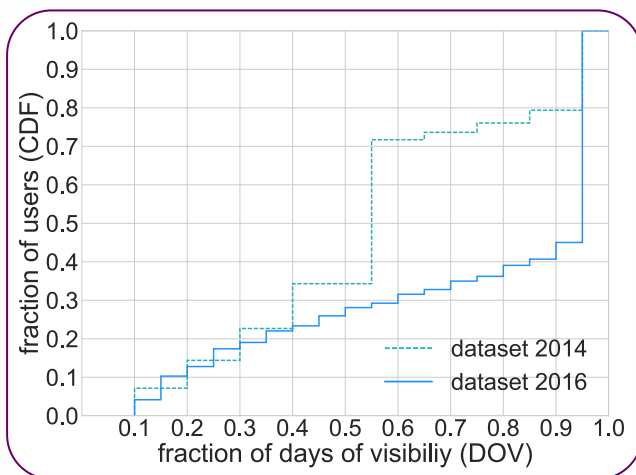
ENTROPY (H)

measure of the uniformity of the distribution of actions over time (i.e., how much users tend to be active for longer periods)

the higher the better (but not too much...)

Data quality indicators

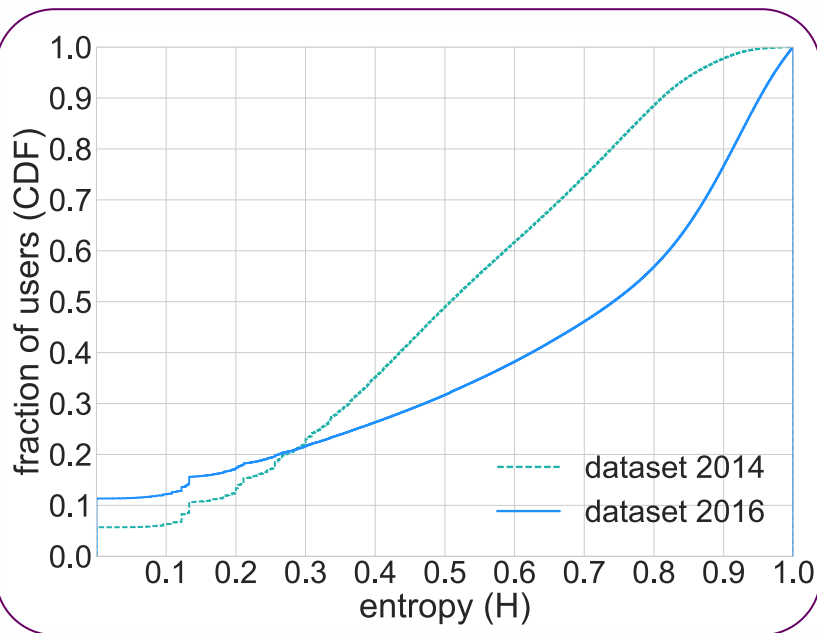
Cumulative distributions across users: **DS2014** vs **DS2016**



Overall improvements of all indicators from 2014 to 2016

Data quality indicators

Comparing entropy of users in DS2014 and DS2016



Entropy: measure of uniformity of actions over time (it summarizes the other indicators)

$$H(X) = \sum_{i=1}^n p(x_i) \log p(x_i)$$

X = distribution of actions over time
 n = number of time slots (e.g., 1-hour)
 x_i = i -th time slot
 $p(x_i)$ = fraction of actions in x_i

Extreme cases:

$H(X_{\text{user a}}) = 0 \Rightarrow$ all actions in one time-bin

$H(X_{\text{user b}}) = 1 \Rightarrow$ constant number of actions

The higher the better: a more uniform behavior allows better tracking users' position

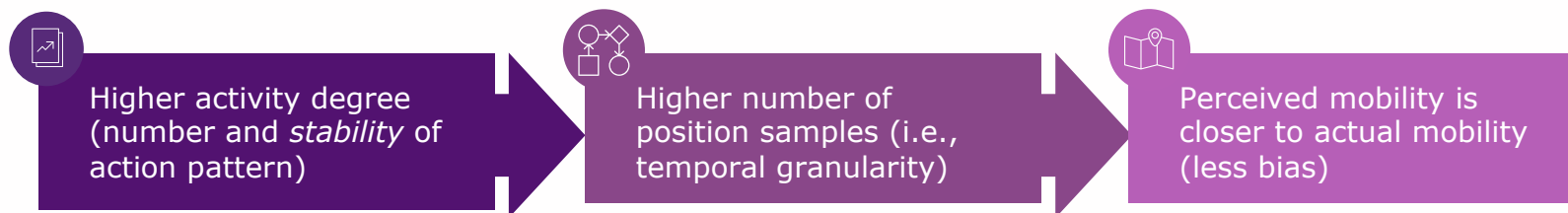
However, $H(X)=1$ is a sign of non-human activity (e.g., IoT/m2m traffic, which has to be excluded)

As for all other indicators, ds2016 is better than ds2014 wrt. users' entropies

Users in ds2016 tend to have higher entropy, which means better overall quality of perceived movements

Highly Active Users

Selecting user samples for human mobility studies



To reduce bias in mobility results, we only consider Highly Active Users (HAU)

Define HAU samples according to requirements

e.g. urban mobility studies requires higher time granularity than Nation-scale studies

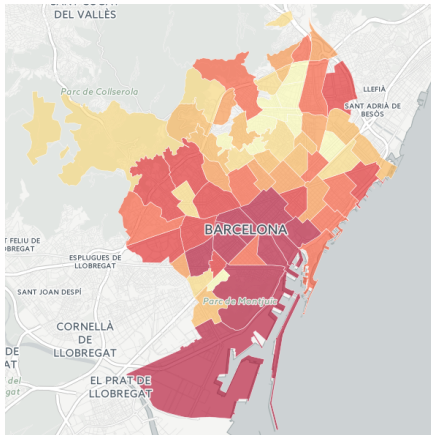
Even with strict requirements, the HAU sample will be large and statistical relevant

users are in general more active and the usage of data connections is widespread

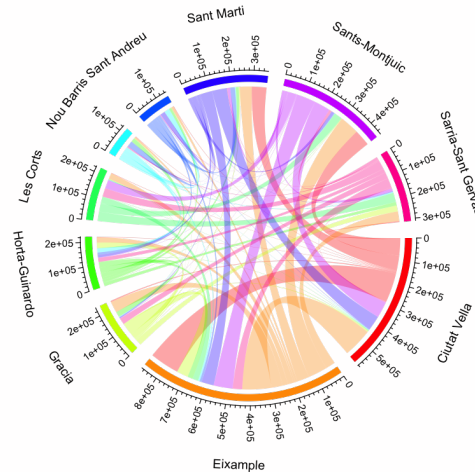
Thresholds from an operational study		ALL conditions	Highly Active Users samples	
Days Of Visibility (DOV)	$\geq 75\%$		DS2014	7.5% (\Rightarrow 25% day only)
Hourly Action Rate (HAR)	≥ 1		DS2016	12.5% (\Rightarrow 38% day only)
Average Lag Time (ALT)	$\leq 30m$		Sample size rather large: up to 38% of entire user population in DS2016	
Total Inactive Time (TIT)	$\leq 75\%$			
Entropy (H)	$0.7 < 0.9$		Statistical relevant samples: demographic (e.g., gender, age) statistical characteristics are preserved	

Applications

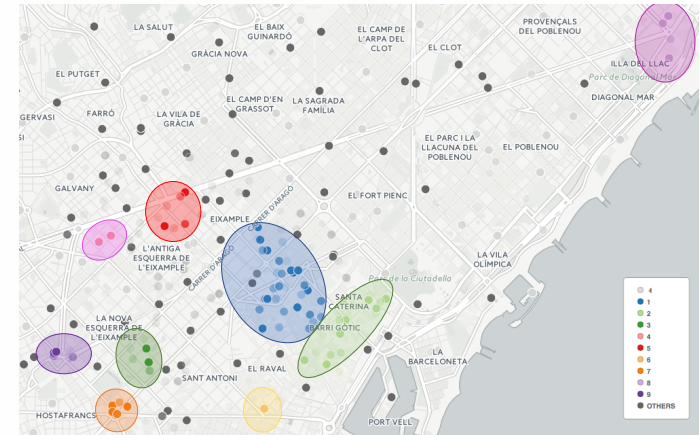
Experience from a commercial project (tourist mobility in Barcelona)



Heat maps: study concentrations of people per area







Transitions: between pairs of PoI, cities or neighborhoods



Human activity clustering: DBSCAN on weighted (by action count) tower locations

Note: all studies are based on HAU sample, in order to reduce the mobility bias introduced by inactivity of users (even for studying static behaviors...)

Conclusions

-  **Longitudinal study of two Nation-wide CDR datasets from 2014 and 2016**
collected in collaboration with major operator
-  **Definition of indicators for describing quality of mobility information**
focus on temporal granularity (number of actions per unit of time and uniformity of activity patterns)
-  **Drastic change in the characteristics of CDR time-granularity over time**
steep trend in the adoption of data connections and increase of recorded action frequency/stability
-  **Interesting outlook for future exploitation of CDRs**
CDR-based studies and products/commercial studies are gaining relevance

GRÀCIES THANK YOU



Speaker: **Alessandro D'Alconzo** alessandro.dalconzo@ait.ac.at
Questions: **Pierdomenico Fiadino** pierdomenico.fiadino@eurecat.org