Big Data for the Social Sciences

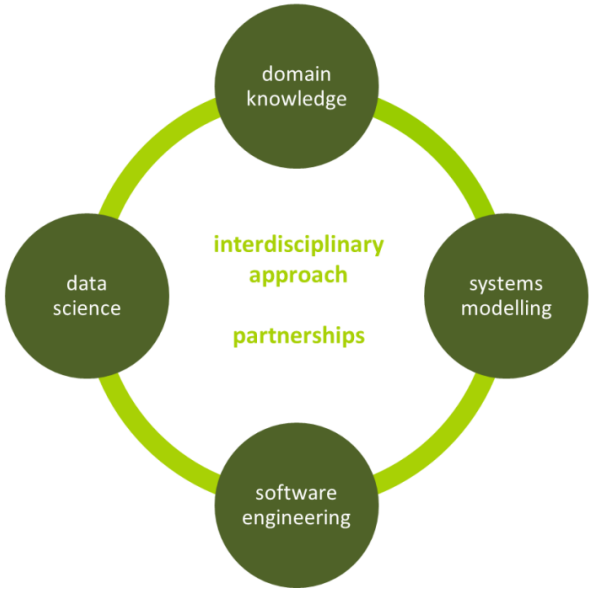
The Example of Mobile Phone Records

1st September 2016

About us

NOMMON

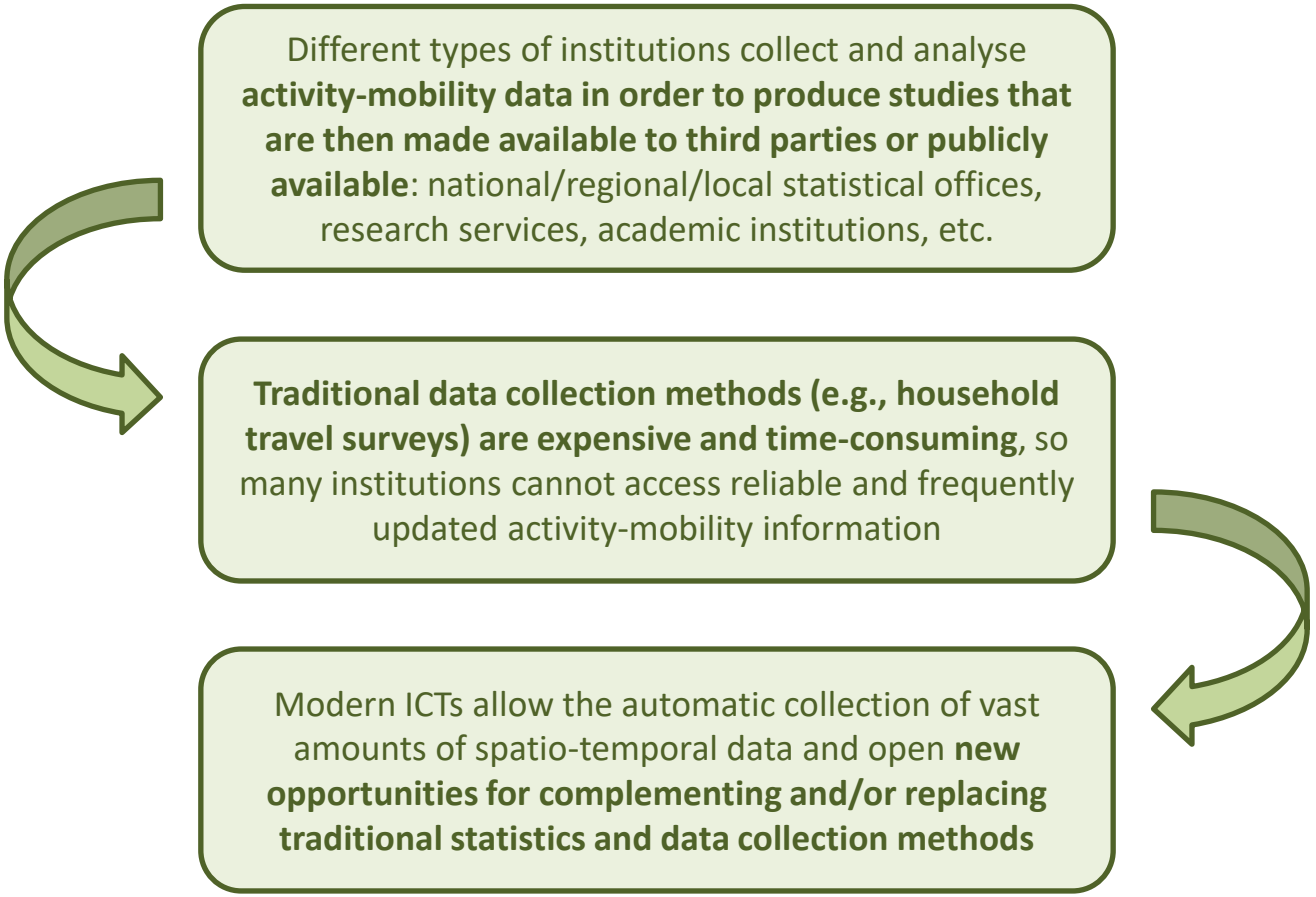
- Technology company based in Madrid
- Decision support tools for the planning and management of transport systems



KINEO
MOBILITY ANALYTICS SL

- Joint company Nommon - Luis Willumsen
- Analysis of geolocated data
- Sectors: transport, tourism, geomarketing…
- Projects in Europe & LatAm

Activity-mobility statistics



Information needs

Distribution of population

Distribution of workplaces

Urban mobility

Interurban mobility

Impact of events

Tourism flows

Data collection: traditional approach

- Surveys:
  - Official statistics
  - Ad-hoc surveys (e.g., mobility surveys)
- Administrative registers
  - Trend towards replacing surveys by administrative registers, where possible

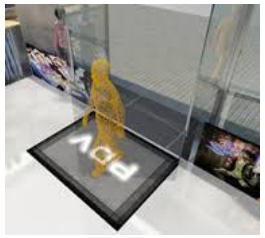
Problems

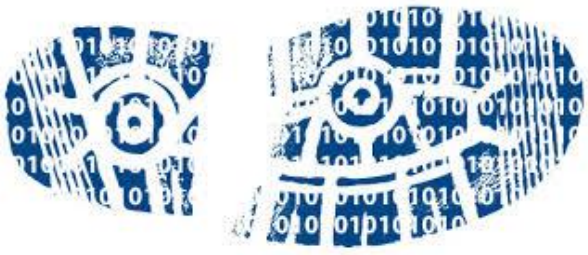
- Information not fully adapted to the problem
- Outdated information
- High cost
- Time consuming
- Small samples
- Quality of information

Opportunities

Smart personal devices

Sensors





What we do

We mine and analyse anonymised geolocation data from mobile devices and blend it with other data provided by our customers or available from public databases **to provide continuously updated and rich activity-travel information in a fraction of the time and cost required by traditional methods**

What we do

Anonymised mobile phone records

Data from other mobile devices

Customer data

Other data

Kineo Cloud Analytics

- Pre-processing & storage
- Data cleansing
- Extraction of activity and mobility patterns
- Expansion to total population

Dynamic population mapping and activity statistics

Trip matrices and mobility statistics

Where are the people?

What are they doing?

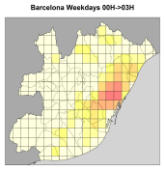
How do they move?

Why?

Dynamic Population Mapping and Activity Indicators

Origin-Destination Matrices and Mobility Statistics

Origin-Destination Matrices and Mobility Statistics



Dynamic Population Mapping and Activity Statistics

Example applications

Transportation

Urban planning

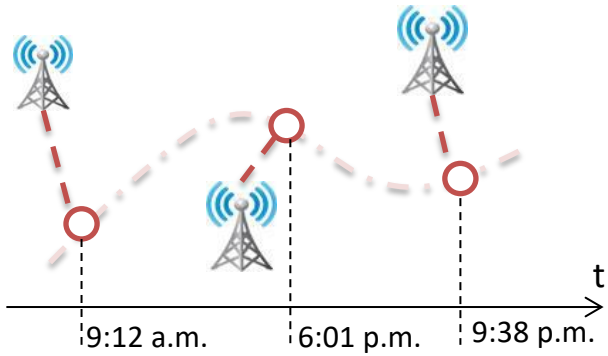
Emergency & disaster management

Tourism

Events

Mobile phone data: what do we mean exactly?

- Typically anonymised call detail records (CDRs) or data from network probes
  - Spatio-temporal data: time and user position every time an event occurs
  - Position typically at cell level (but some MNOs are beginning to store triangulated positions)
- MNO's clients + roamers that connect to its network
- Sociodemographic data for clients (age and gender)



So how does it work?

What we do (more or less):

- Sample construction
- Identification of users' home
- Identification of other activities (work, study…)
- Generation of activity-travel diaries
- Expansion to the total population
- Calculation of indicators

Why mobile phone records?

Great potential for the analysis of human activity and mobility:

- Passive data collection eliminates many of the intrinsic limitations of surveys, such as incorrect and imprecise answers
- Opportunistic data collection – no need for specific infrastructure
- Reasonably good temporal and spatial granularity
- Large samples
- Longitudinal data
- Opportunities for “natural experiments”
- Drastic cost and time reduction with respect to traditional methods

But…

These data also come with a number of challenges:

- Noise and errors
- Biases
- Less explanatory power
- Computational challenges

$\Rightarrow$ Mobile phone data do not provide the full picture: need for data fusion

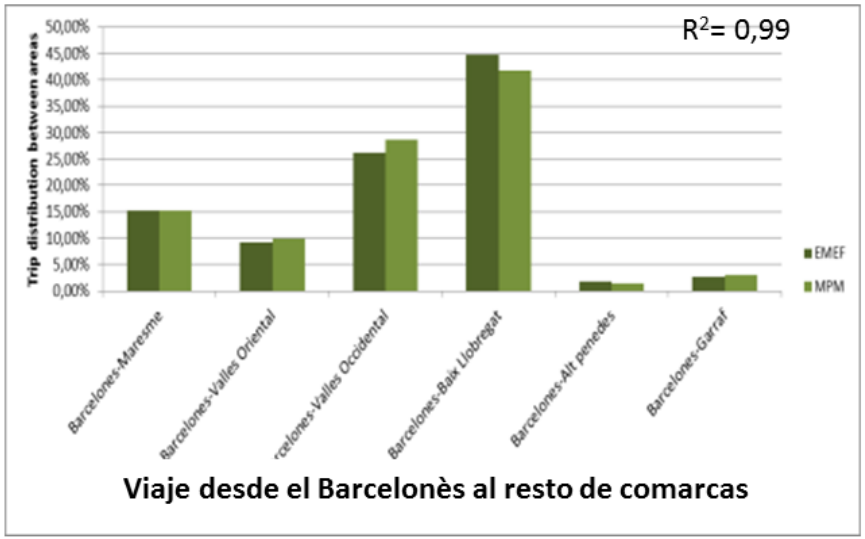
$\Rightarrow$ Algorithms matter: devil's in the details

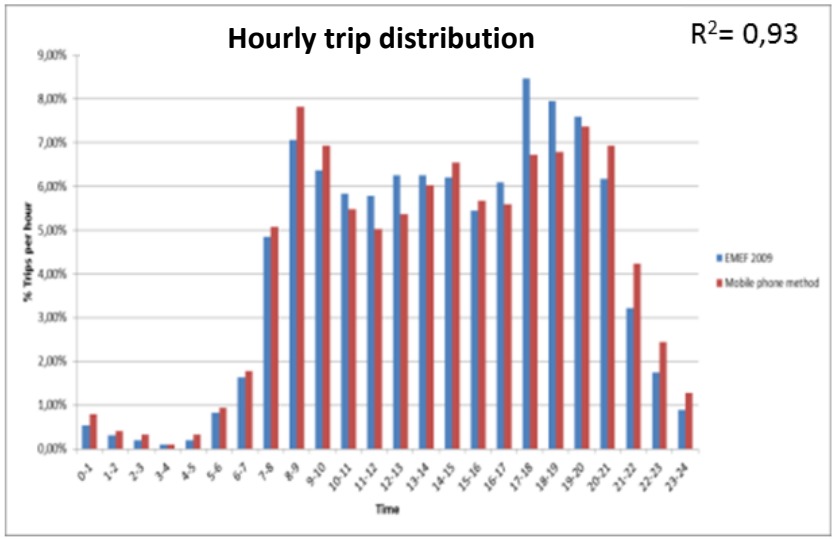
$\Rightarrow$ Rigorous validation is essential

Example applications

OD Matrices for Barcelona Metropolitan Area

- Calculation of origin-destination matrices in the Barcelona Metropolitan Area and comparison with the Mobility Survey on Working Days (EMEF)



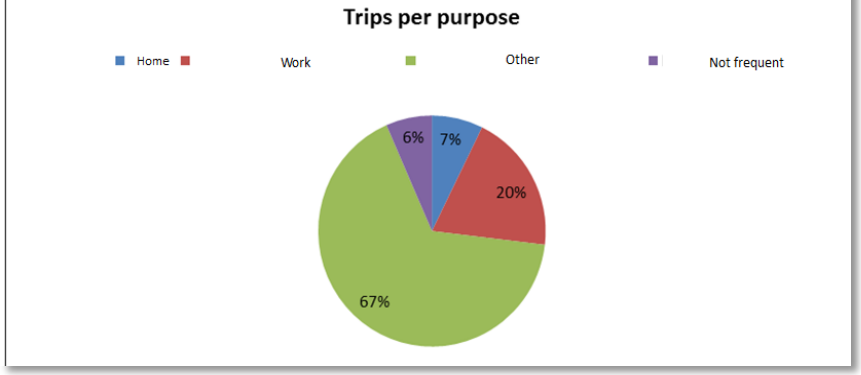


eunoia EVOLUTIVE USER CENTRIC NETWORKS FOR INTRAURBAN ACCESSIBILITY

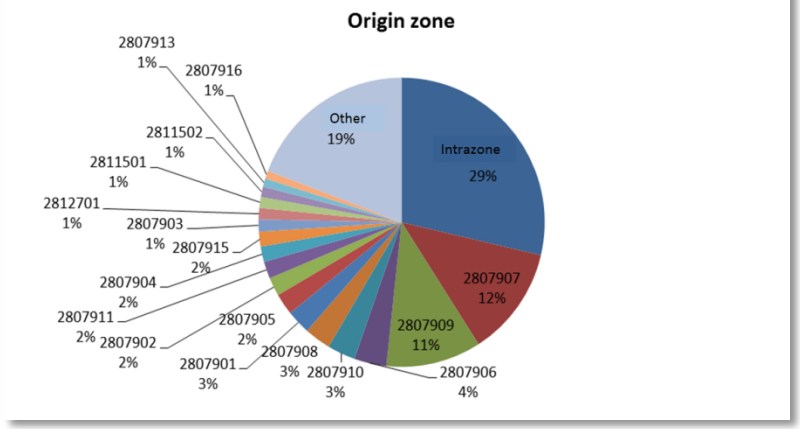
Trip Attraction Areas in Madrid

- Trip matrices for several attraction areas in the Region of Madrid

Study area	Ciudad Universitaria (zone_1)			
Trips	358.396			
Trip Purpose	**Home**	**Work**	**Other**	**Not frequent**
Trips per purpose	26.049	70.315	238.554	23.478

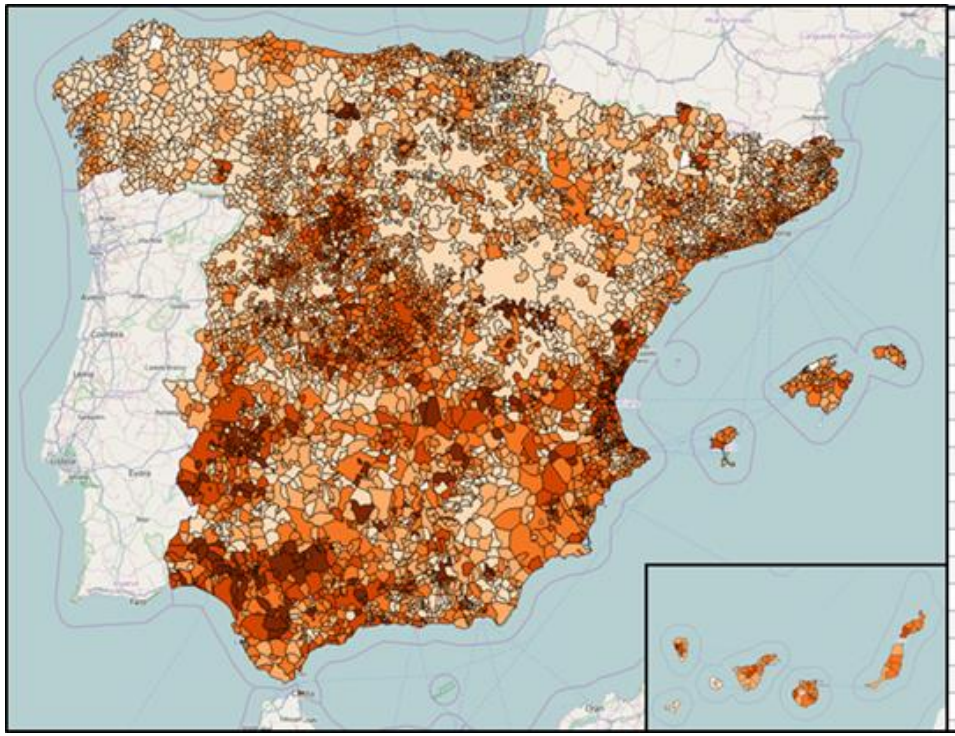


Number of origin zones	239/307
Origin zones that account for the 80% of the trips	18



Analysis of Commuting Patterns in Spain

- Commuting Patterns in Spain: pilot study for next Spanish Census

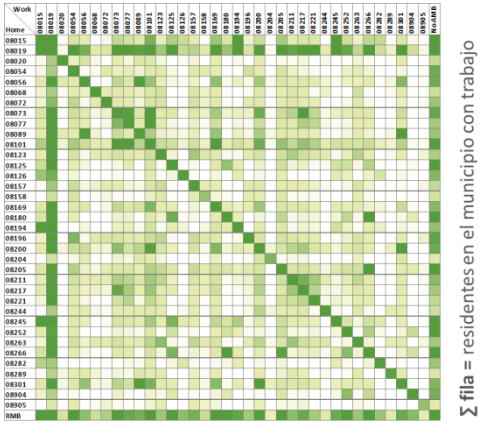


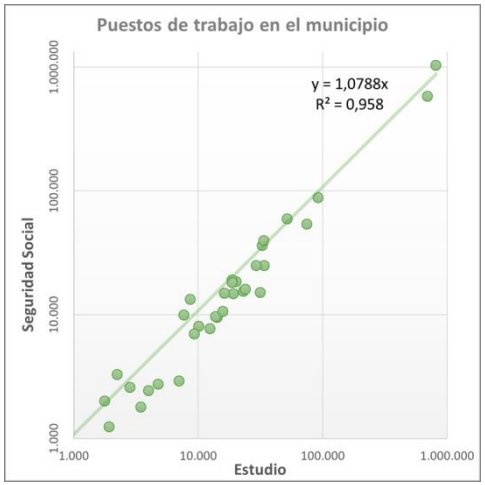
Sample	Sample %	ID	Population
111	7.602739726027398	0104001	1460.000000000000000
80	5.925925925925927	0104002	1350.000000000000000
2834	11.626666666666667	0104003	24375.000000000000000
23	3.357664233576643	0104005	685.000000000000000
1114	10.257826887661142	0104006	10860.000000000000000
36	3.956043956043956	0104007	910.000000000000000
35	5.600000000000001	0104008	625.000000000000000
14	2.043795620437956	0104010	685.000000000000000
160	4.221635883905013	0104011	3790.000000000000000
23007	12.174948404508653	0104013	188970.000000000000...
33	17.368421052631579	0104014	190.000000000000000
365	11.196319018404909	0104016	3260.000000000000000
166	3.660418963616317	0104017	4535.000000000000000
33	9.565217391304348	0104018	345.000000000000000
16	4.266666666666667	0104020	375.000000000000000
22	9.777777777777779	0104021	225.000000000000000
12	1.276595744680851	0104022	940.000000000000000
536	12.807646356033453	0104024	4185.000000000000000
23	8.363636363636363	0104028	275.000000000000000
1710	11.350813143046798	0104029	15065.000000000000000
243	16.933797909407666	0104030	1435.000000000000000
406	10.545454545454545	0104031	3850.000000000000000
967	12.279365079365078	0104032	7875.000000000000000
883	6.740458015267175	0104035	13100.000000000000000
21	7.000000000000001	0104036	300.000000000000000

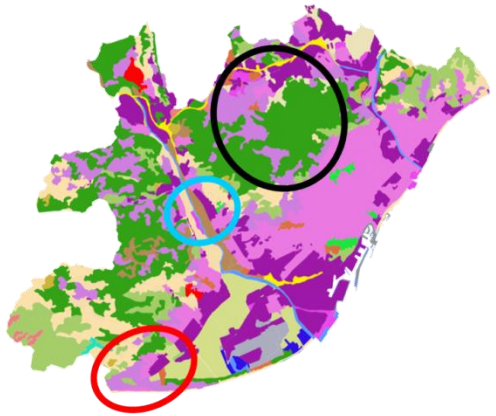
Instituto Nacional de Estadística

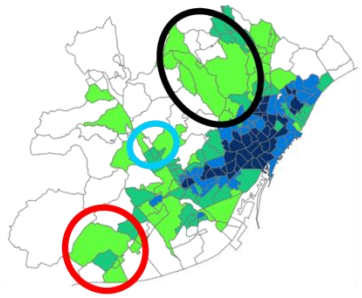
Analysis of Workplaces Distribution in Barcelona

- Home-Work matrices for the Metropolitan Area of Barcelona

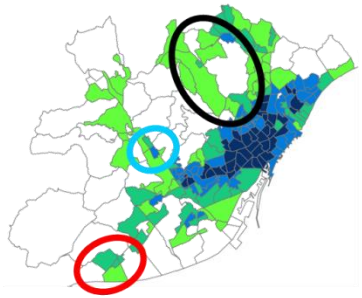








Asignación sin usos del suelo

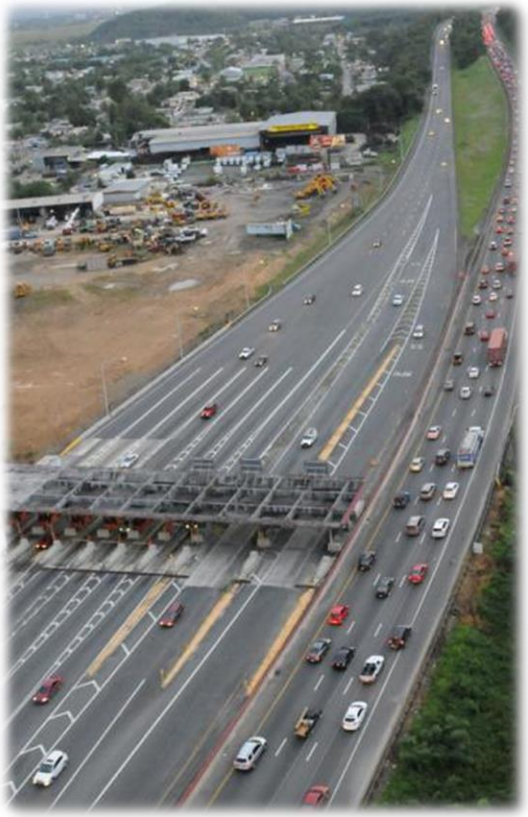


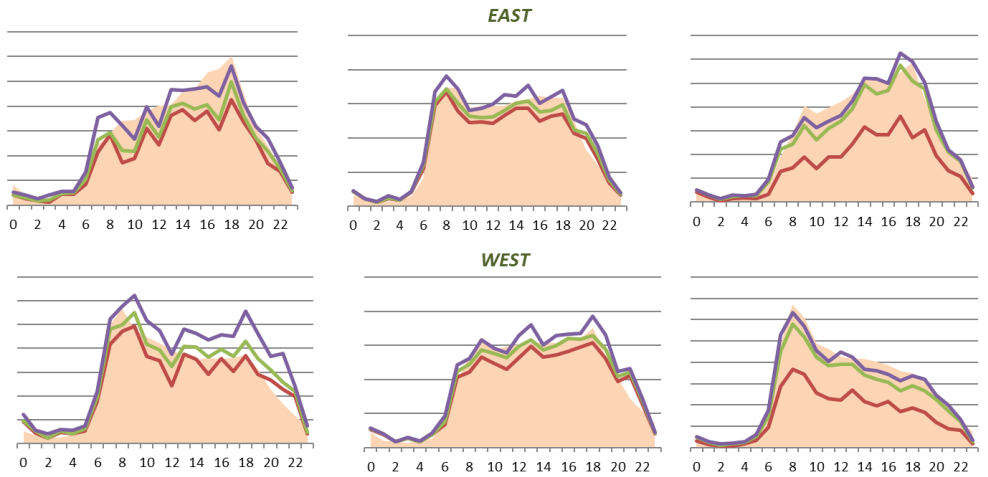
Asignación con usos del suelo

BR BARCELONA REGIONAL AGÈNCIA DESENVOLUPAMENT URBÀ

Analysis of Toll Road Potential Demand

- Estimation of potential demand for a toll road and modelling of different revenue optimisation strategies

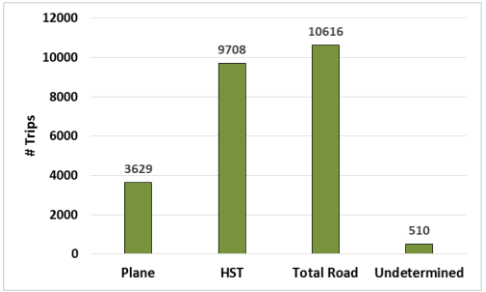


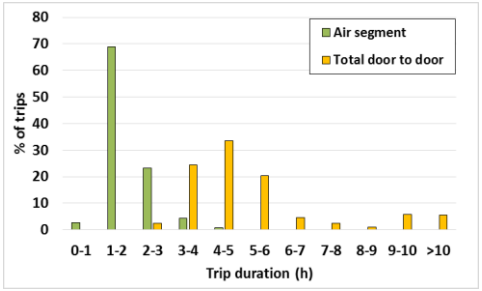


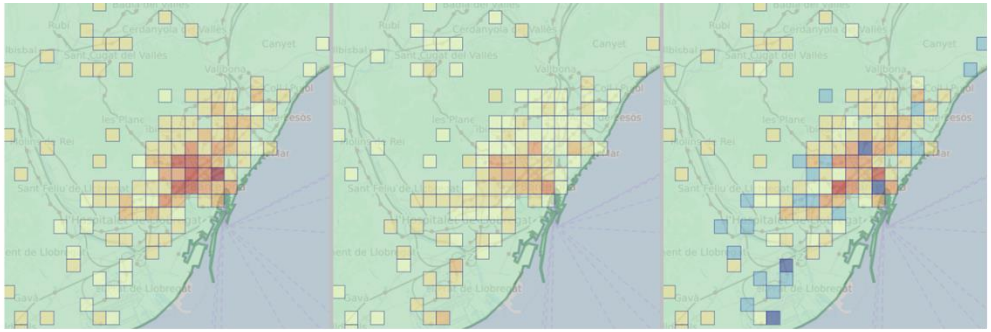
cintra

Analysis of Madrid-Barcelona Transport Corridor

- Analysis of modal split, airport and HST catchment areas, and door-to-door travel times







No Train trips

No Air trips

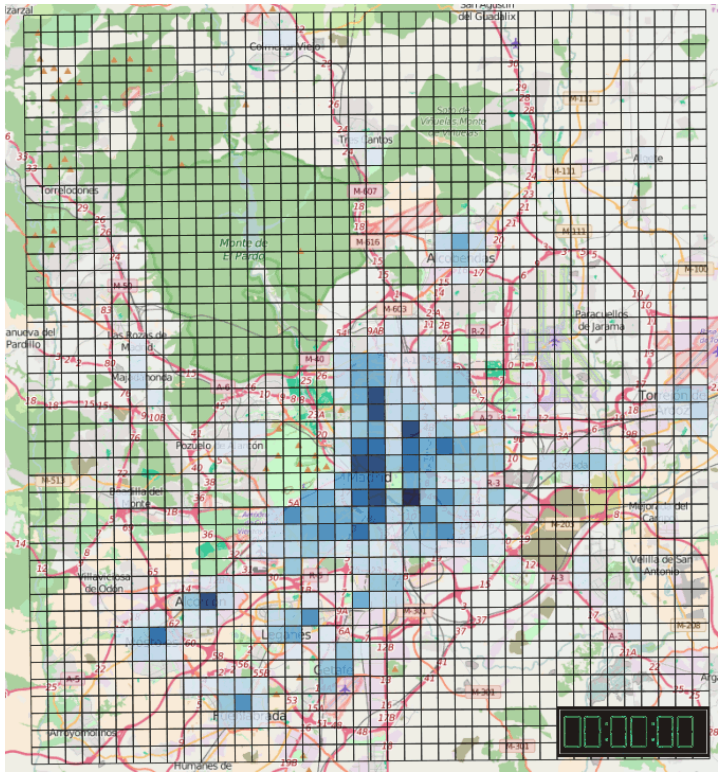
[No Train trips - No Air trips]

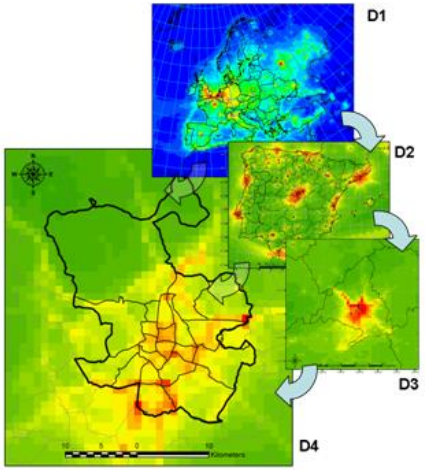
BigData4ATM

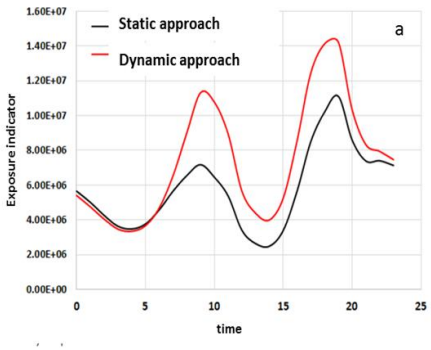
SESAR JOINT UNDERTAKING

Population Exposure to Air Pollution

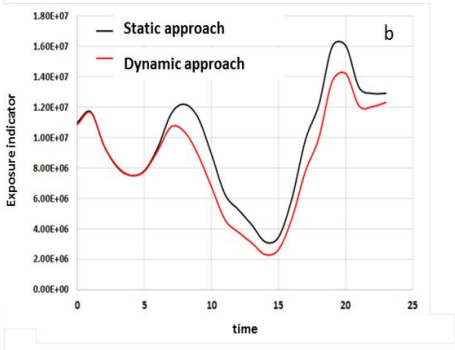
- Exposure of population to NOx emissions in Madrid based on dynamic population mapping







Chamartín



Carabanchel

¡MADRID! ÁREA DE GOBIERNO DE MEDIO AMBIENTE Y MOVILIDAD

Some conclusions…

The promise of Big Data

Big data streams generated by smart personal devices offer a great potential for analysing human activity-mobility patterns and producing sociodemographic statistics

Yet, some precautions…

Data is not (always) a substitute for theory

“Models without data are just stories - data without models are just numbers”

Yet, some precautions…

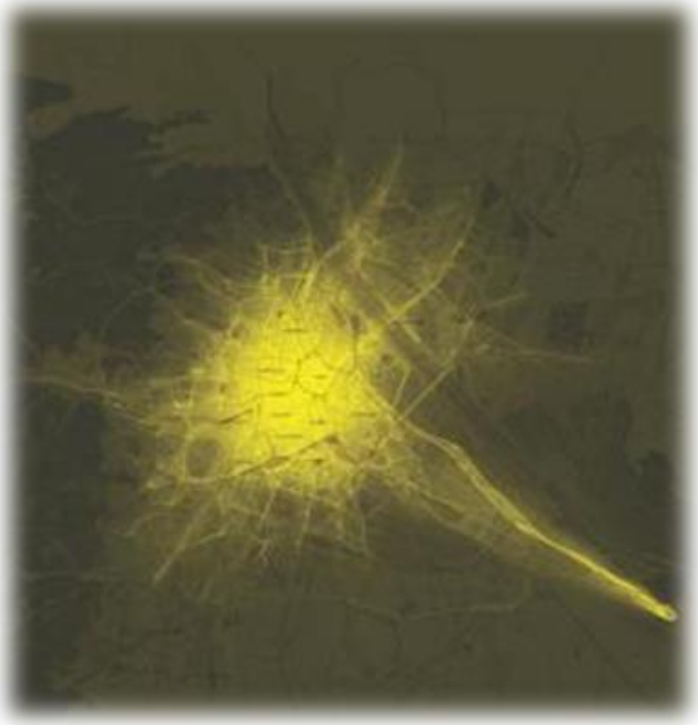
Neither data nor algorithms are “neutral”: lack of understanding of underlying principles and limitations may lead to a false sense of objectivity and democratisation of information

Critical interpretation of the results in terms of the initial problem is essential, and we (researchers, consultants…) must do so responsibly

Yet, some precautions…

Big Data is usually noisy, unstructured and biased (“all data are biased, but some are useful”), so producing actionable information requires a lot of effort for data cleaning, data fusion, and algorithm validation…

…but this is the hard part, not really cool, so many people prefer to focus on fancy (and often useless) visualisation



Yet, some precautions…

Big Data also comes with a number of regulatory, organisational and business challenges

(We won’t discuss this today, but it should not be forgotten)

We have to live with

- Those overselling the benefits of Big Data as the cure for all diseases: more common in industry, but also happens in public administration and academia
  - But we can and must be critical and claims to be rigorously supported
- Those unable to see the added value: equally common in industry, in public administration and in academia, but probably more reproachable in the latter case ("I'm happy with my good old crappy data, I still get my papers published")
  - Nothing much to do here - probably the same people who 20 years ago didn't see the added value of Internet over teletext

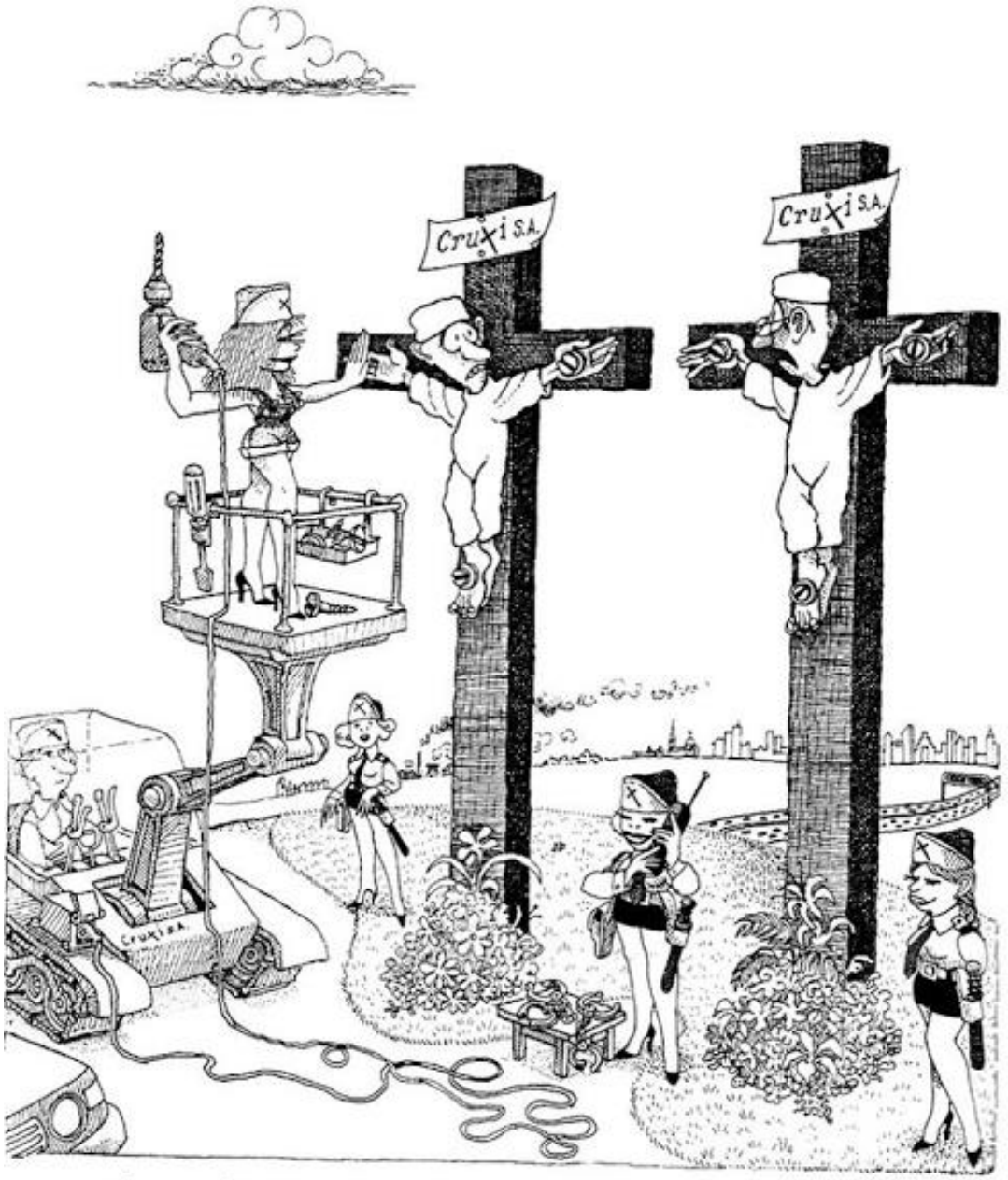
“A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it”

Max Planck

…and one fundamental open question

At present

Big Data has already yielded important practical advances in fields like transportation planning, but so far this is mostly about better/cheaper ways of measuring the same indicators and use them to feed the same models



~ SÍ, EL MÉTODO SIGUE SIENDO ANTICUADO, PERO HAY QUE RECONOCER QUE LA COSA MEJORÓ MUCHO LUEGO DE LA PRIVATIZACIÓN.

In the future

For sure, Big Data will produce new indicators that couldn't be measured before, and will help us make the most of state-of-the-art theoretical frameworks, but…

.. just as the telescope revolutionised astronomy or the microscope revolutionised biology, can Big Data inspire the development of new theories and revolutionise the social sciences?

www.nommon.es

e-mail: nommon@nommon.es

www.kineo-analytics.com

e-mail: kineo@kineo-analytics.com