The European Commission's science and knowledge service

Joint Research Centre

New territorial analyses enabled by emerging sources of geospatial data

Filipe BATISTA, Ricardo BARRANCO, Konstantin ROSINA, Carlo Lavalle
JRC.B.3 – Territorial Development unit

ESPON Scientific Conference 2018
London, UK
14th November 2018

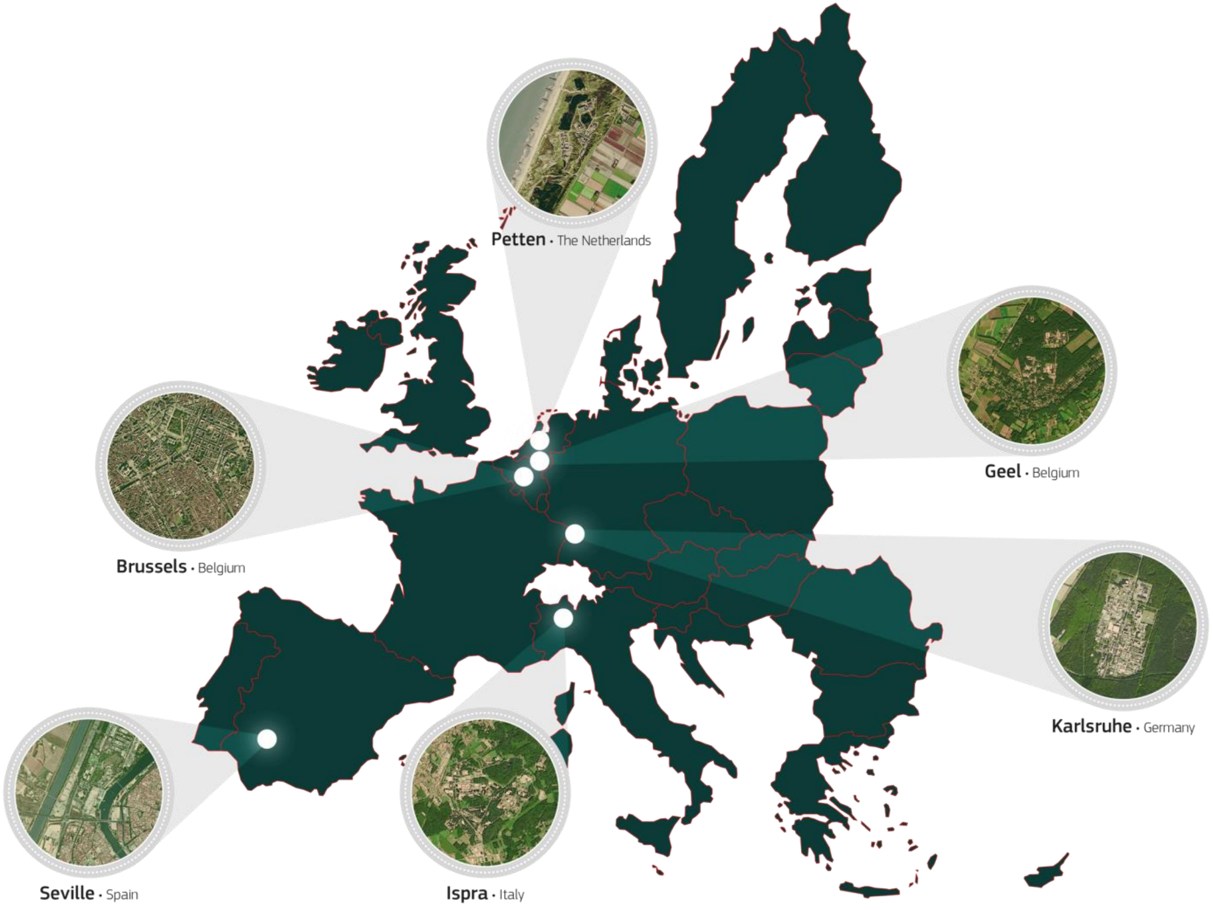
European Commission

Outline

- The Knowledge Centre for Territorial Policies (KCTP)
- Emerging sources of geospatial data
- Potential for urban and regional analyses (examples)
- Discussion and conclusions

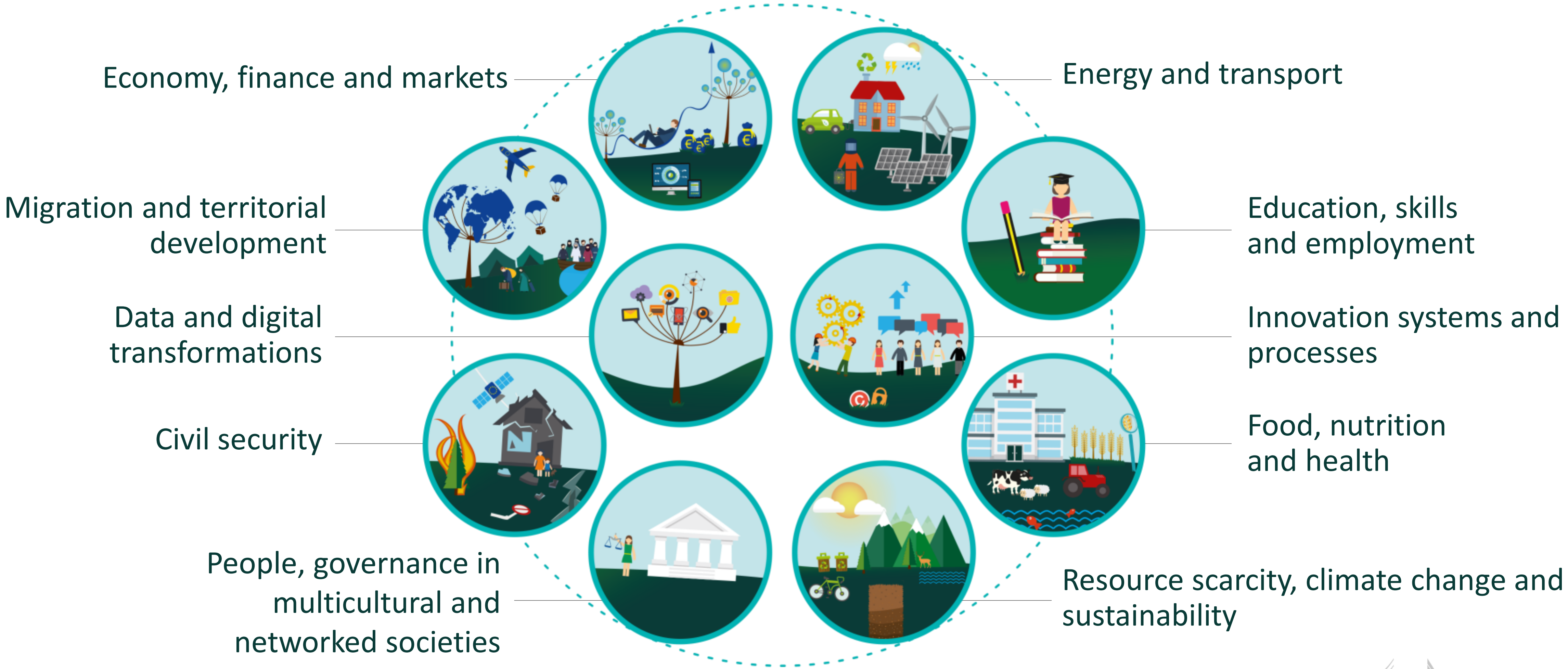
The JRC at a glance

- European Commission's science and knowledge service.
- Supports EU policies with independent scientific evidence.
- 3000 staff (3/4 research staff)
- Headquarters in Brussels + research facilities in 5 Member States
- +1400 scientific publications yearly



European Commission

The JRC at a glance



European Commission

The Knowledge Centre for Territorial Policies

- Part of a wider European Commission strategy on "Knowledge 4 Policy" aiming at improving communication and **interaction between science and policy**.
- The **KCTP** aims at supporting territorial (urban & regional) development policies by promoting better holistic knowledge management and dissemination.

Key components:

✓ **Knowledge base** (data, indicators)

✓ **Analytical and modelling capacity**

✓ Community of Practice on Cities (CoP-Cities)

✓ Field studies (City-labs)

✓ Web platforms

http://ec.europa.eu/knowledge4policy/territorial

European Commission

Emerging geospatial data sources

Conventional inputs for territorial analyses

- Statistical and geographical data from official bodies, surveys, interviews…

Big data paradigm

- ICT-based services generate massive amounts of geo-referenced or geo-tagged data as either final or by-products.
- Data with new characteristics: Volume, Velocity, Variety…
- Can be exploited for new territorial analyses.
- Applications growing at a fast pace also in the geospatial and urban/regional domains.

European Commission

Emerging geospatial data sources

Data generated as a by-product

Unintentional crowd-sourced data.

- Mobile network operator (MNO) data
- Web activity (content, traffic, searches…)
- Social media (tweets, check-ins, photos…)
- Transactions (costumer, financial…)

Data generated on purpose

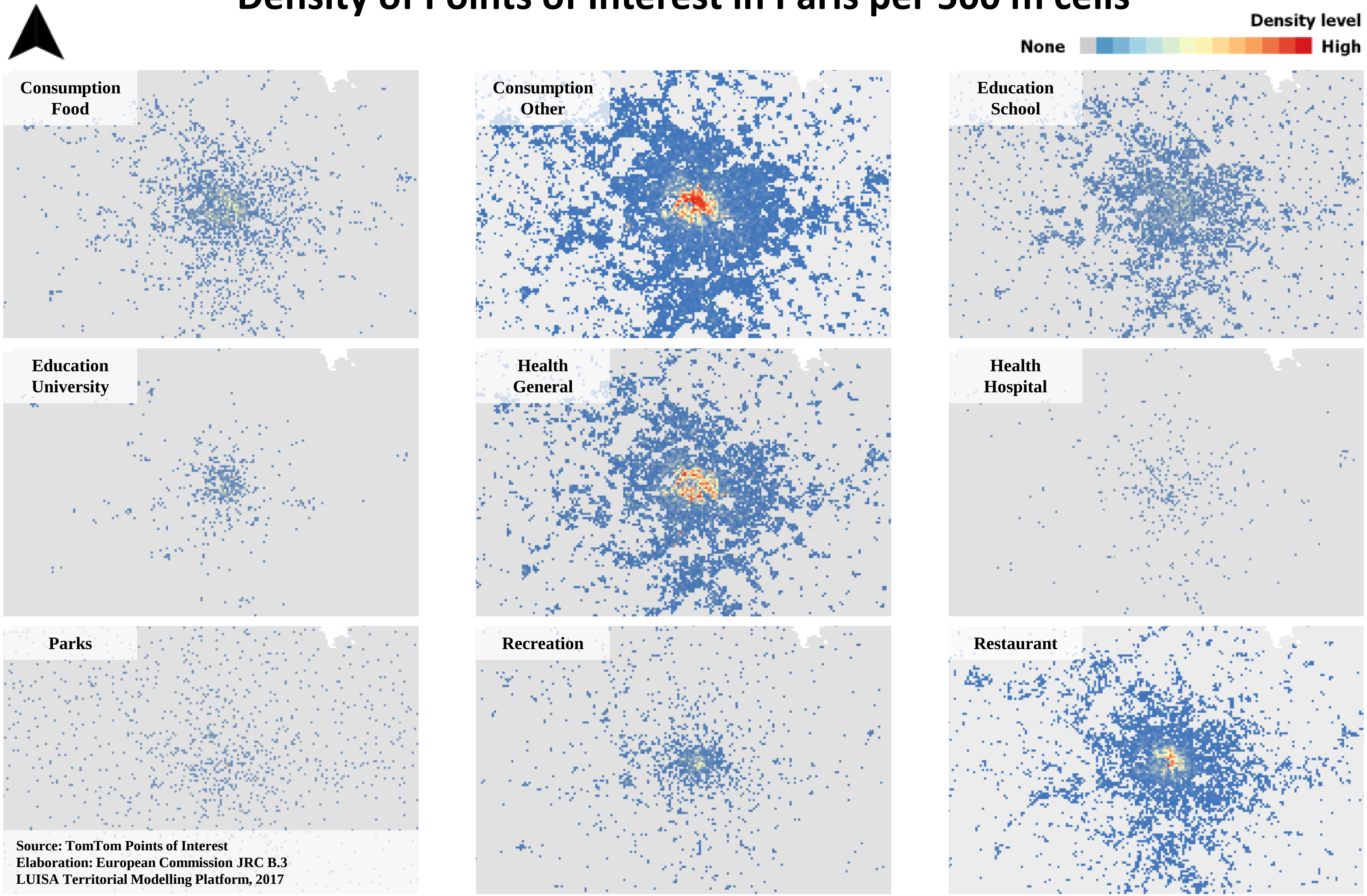
Intentionally produced as a core component of ICT-based service. Aspects in common with of Big Data.

- Navigation/mapping data (e.g. POI, road networks), but volunteered/collaborative or private initiatives
- Sensors (count of vehicles, pedestrians, air-borne, satellite, weather stations)

Points of Interest (POI)

- Physical structures on the Earth’s surface with a functionality relevant to human or societal activities.
- Mapped as a precise points on a (digital) map.
- Many sources:
  - OpenStreetMap (VGI, free and open source)
  - Navigation /mapping / sector data (proprietary) (e.g. TomTom)
  - Derived from mining web services (e.g. Booking.com, TripAdvisor)
  - Different levels of quality, completeness, overlap
  - Different classification systems, semantics and ontologies
  - Different quality (completeness, accuracy…)

Density of Points of Interest in Paris per 500 m cells

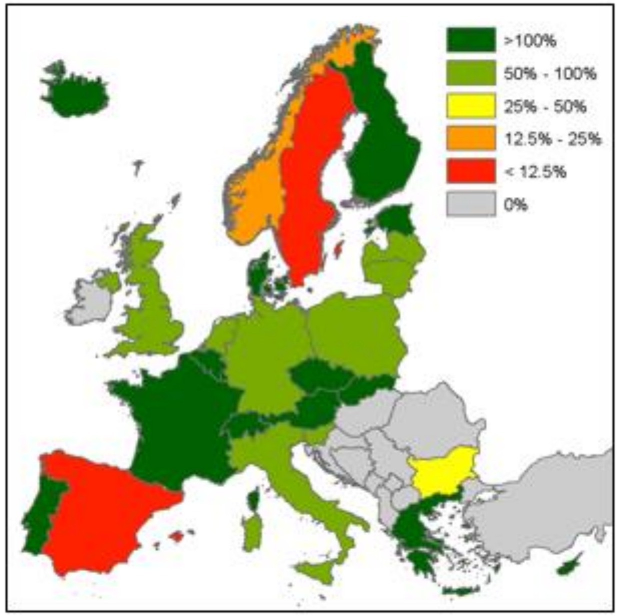


POI data – Quality issues

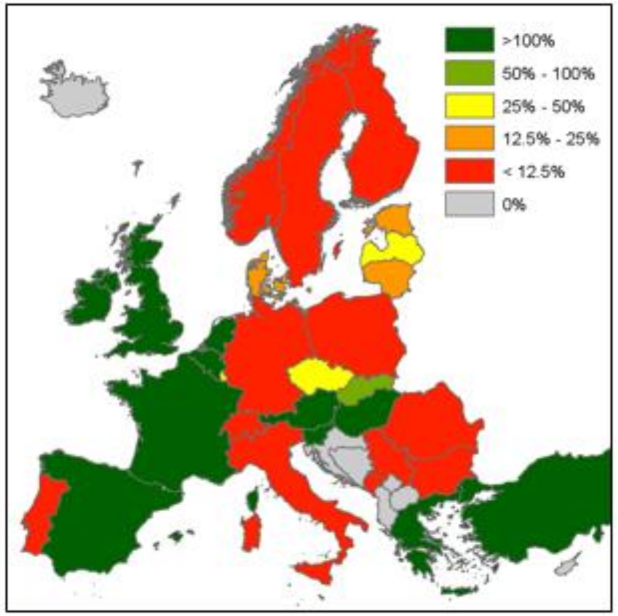
Completeness levels difficult to assess objectively. Looking at available POIs per capita is a possible indirect way to assess completeness.

Number of Education-related POIs per inhabitant as percentage of European average

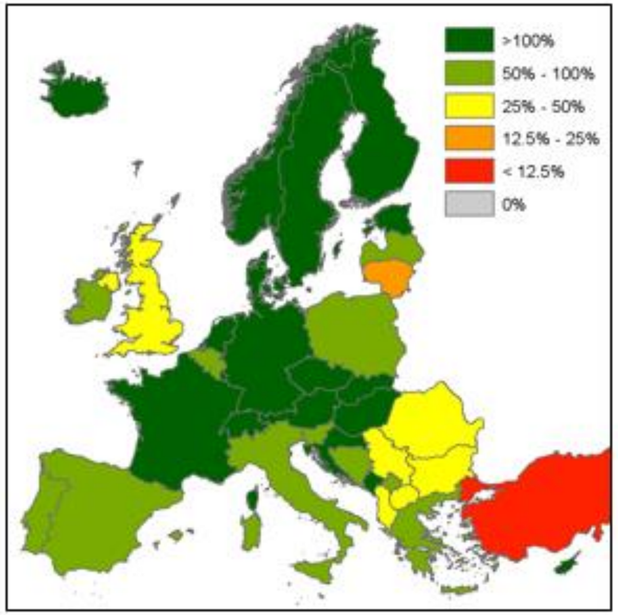
EuroRegionalMap 9.1



TomTom (2014)



OSM (2017)



POI data – Application

Land use characterization using POI data

Part of a wider project to refine the thematic and spatial detail of CORINE Land Cover and map spatiotemporal population densities (ENACT).

Main objective: To break down CLC class 121 (“Industrial and Commercial Sites”) into 3 more detailed land uses.

CLC1	CLC2	Level 2 Label	CLC3	Level 3 Label	CLC4	Level 4 Label
1 Artificial surfaces	11	Urban fabric	111	Continuous urban fabric	1111	Urban fabric dense
			112	Discontinuous urban fabric	1121	Urban fabric medium density
					1122	Urban fabric low density
					1123	Urban fabric very low density / isolated
	12	Industrial, commercial and transport units	121	Industrial and commercial units	1211	Production facilities
					1212	Commercial / service facilities
					1213	Public facilities
			122	Road or rail networks and associated land	1221	Road / rail networks and associated land
					1222	Major stations
			123	Port areas	1231	Port areas
			124	Airport areas	1241	Airport areas
					1242	Airport terminals
	13	Mines, dumps and construction sites	131	Mineral extraction sites	1311	Mineral extraction sites
			132	Dump sites	1321	Dump sites
			133	Construction sites	1331	Construction sites
	14	Artificial vegetated non-agricultural areas	141	Green urban areas	1411	Green urban areas
			142	Sport and leisure facilities	1421	Sport and leisure green
					1422	Sport, leisure and touristic built-up

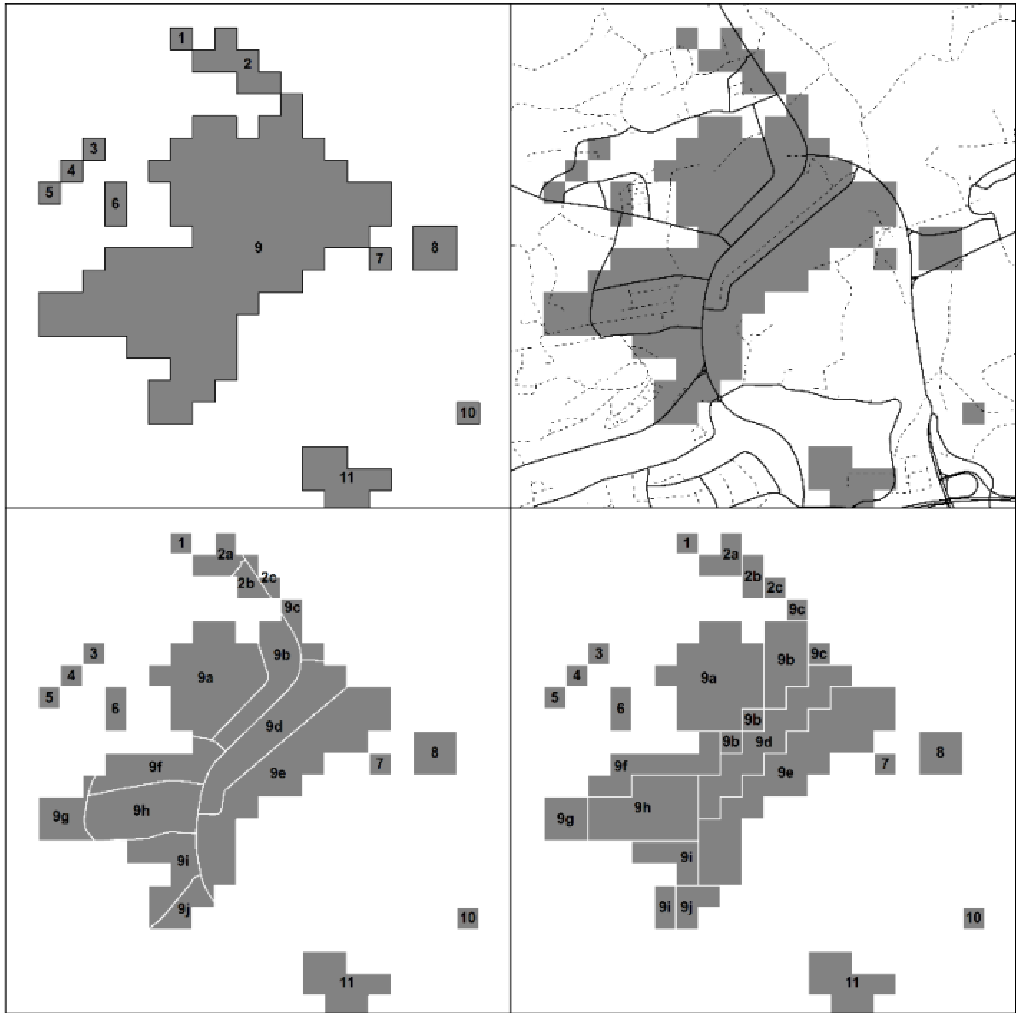
POI data – Application

Land use characterization using POI data

Main steps:

1. Segmentation of industry-commerce-service clusters
2. Labelling of the training set
3. Feature engineering
4. Supervised machine learning classification
5. Validation

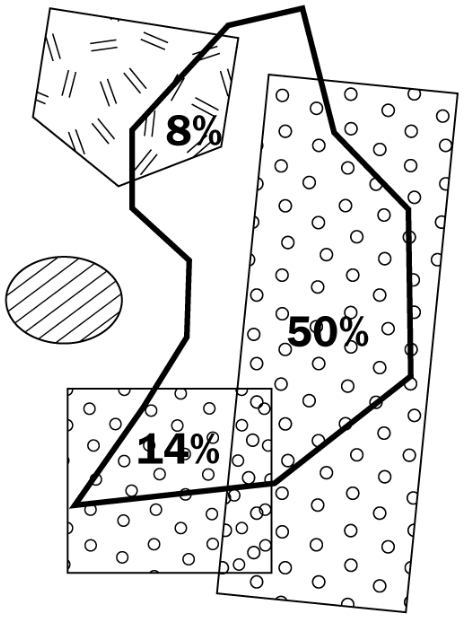
Segmentation of IC clusters by road network



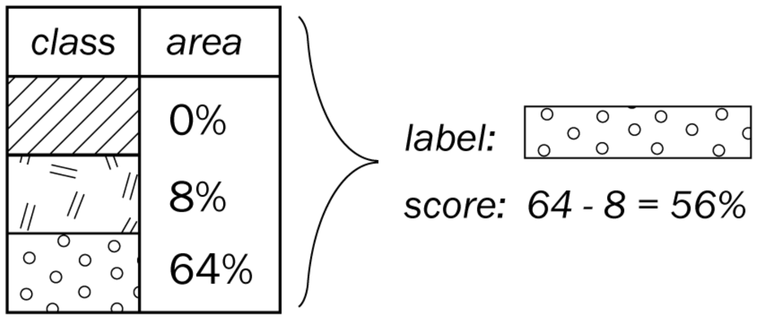
- Contiguous pixel clusters
- Relevant road categories from TomTom
- Over 700,000 segments

Labelling of the training set

- Overlay with three sources of reference land use polygons
- TomTom, OpenStreetMap, National/regional databases
- 270,000 labelled segments (with score >30%)

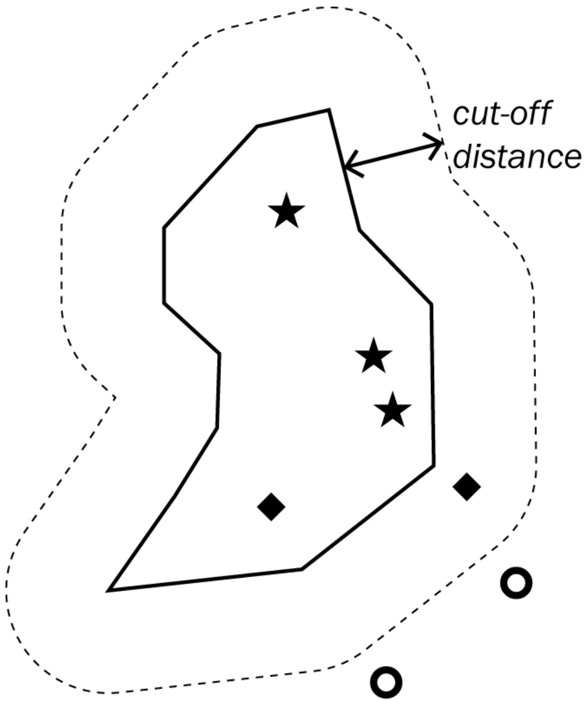


Cummulative relative intersection area per class, across datasets



Feature engineering

- Mining of spatial context from available geodata
- Primary context – data directly related to economic activity
  12 million POIs (TomTom, OpenStreetMap, EuroRegionalMap...)



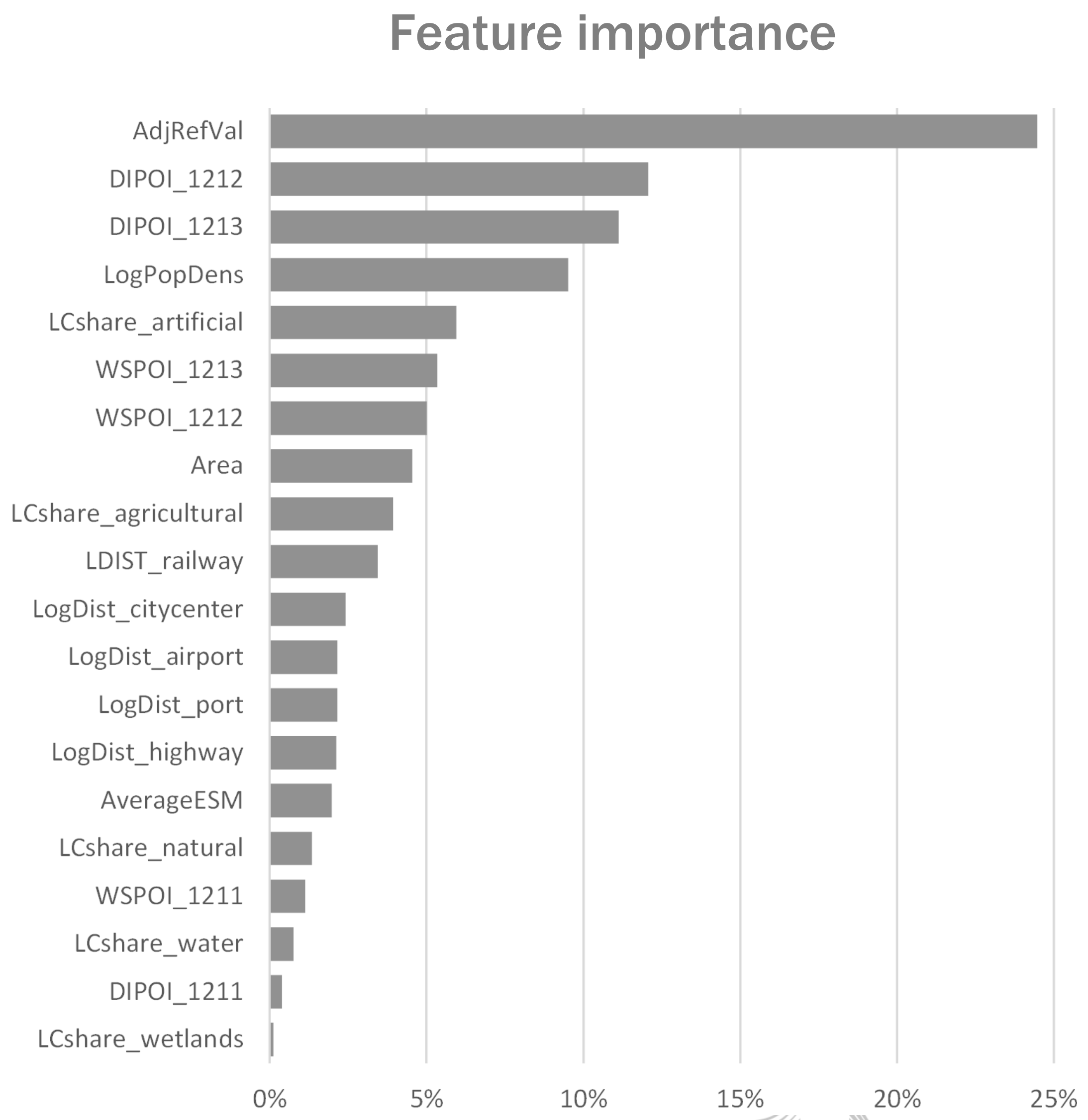
Distance-weighted sum of POI per category

categ.	score
○	0.00
◆	1.25
★	3.00

- Secondary context:
  adjacent label, built-up/ population density, LULC composition, proximity to highway, railway, airport, city centre

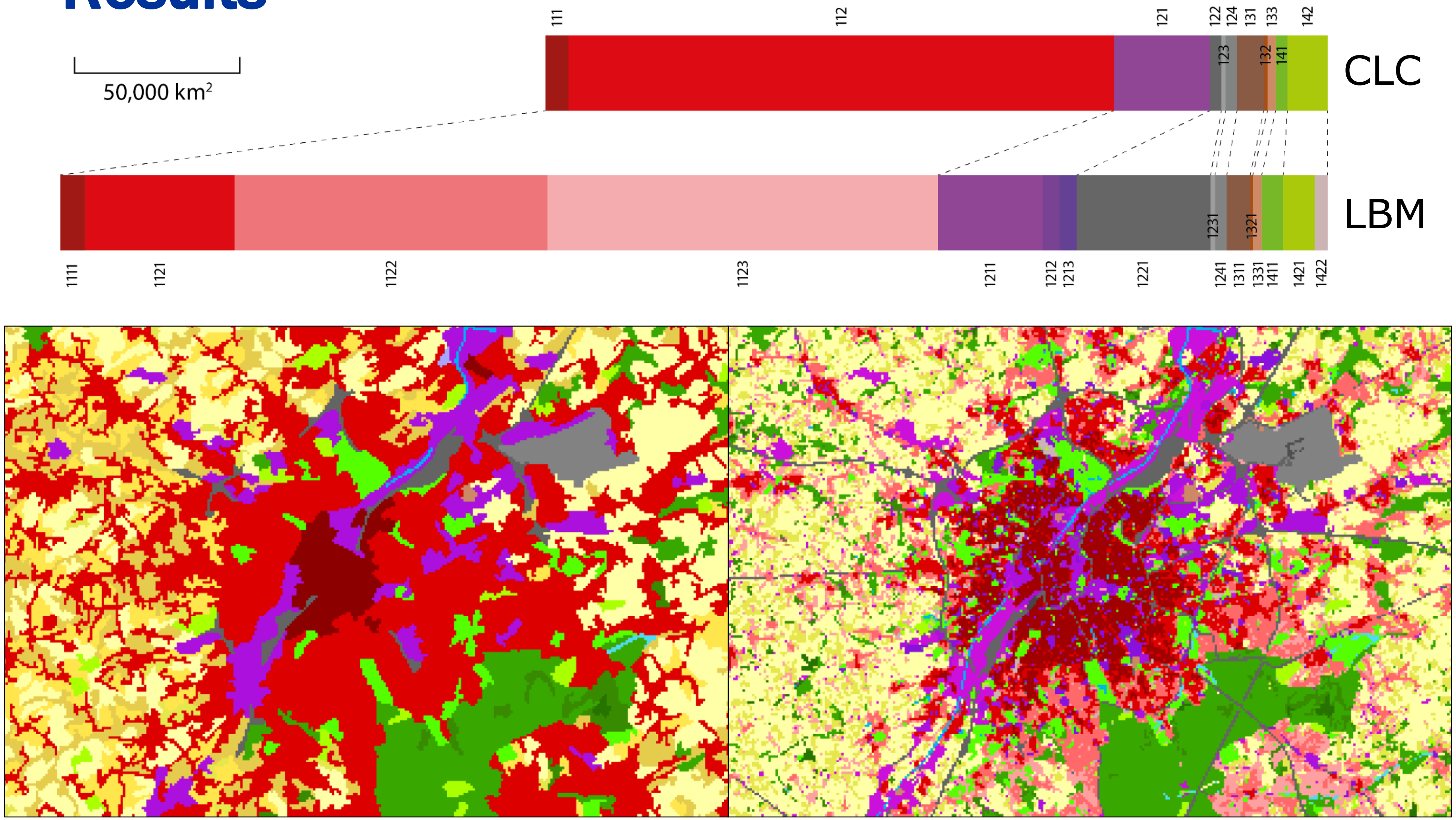
Classification

- Random forests classification with H2O library + R
- 50 models run
- Several specifications of the training set (controlled by various label scores),
- Several selections of predictor variables
- Transformations of predictor variables
- The best model:
  - Training set reduced to label scores above 50 (ca 225,000 polygons, 30% was left aside for testing)
  - Ensemble of 1000 trees was used
  - Log transformed proximity and density variables



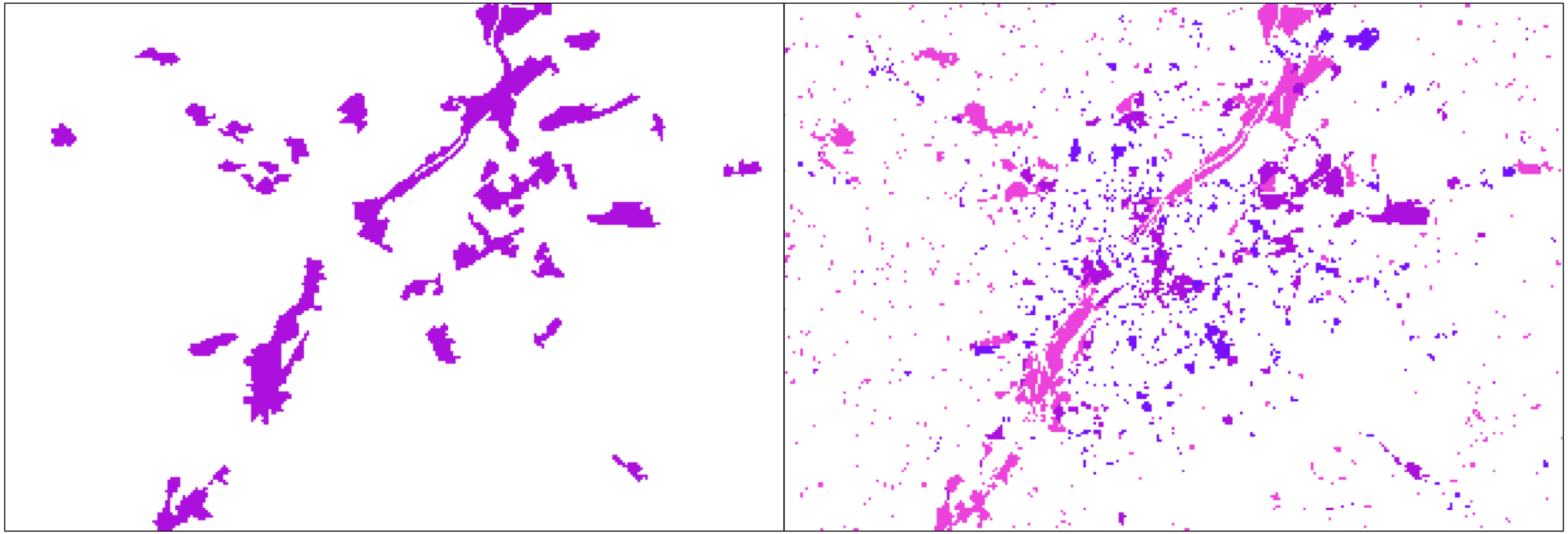
European Commission

Results



50,000 km^2
111
112
121
122
124
131
133
142
123
132
141
CLC
LBM
1231
1321
1111
1121
1122
1123
1211
1212
1213
1221
1241
1311
1331
1411
1421
1422

Results



121 Industrial and commercial units

1211 Production facilities (ABCDE)
1212 Commercial/service facilities (GHIJKLMN)
1213 Public facilities (OPQ)

European Commission

Accuracy assessment

Machine learning model performance

Overall accuracy: 87.4% Cohen's Kappa: 0.72		Prediction				
		1211	1212	1213	Total reference	Omission error
Reference	1211	44,106	1,262	870	46,238	4.6%
	1212	3,298	6,831	558	10,687	36.1%
	1213	2,229	285	8,166	10,680	23.5%
	Total predicted	49,633	8,378	9,594	67,605	
	Commission error	11.1%	18.5%	14.9%		

Independent validation of the final map

Overall accuracy: 74.0% Cohen's Kappa: 0.53		Prediction				
		1211	1212	1213	Total reference	Omission error
Reference	1211	232	16	9	257	9.7%
	1212	73	40	5	64	66.6%
	1213	13	5	73	61	19.8%
	Total predicted	318	61	87	466	
	Commission error	27.0%	34.4%	16.1%		

European Commission

Web content mining

Applied to **extract useful information from websites**.

Many applications:

Non directly geospatial-oriented

- Media monitoring (e.g. EMM).
- Mining of prices for price indexes, inflation rates.
- Citizen and costumer sentiment (widely used by private sector to optimize business).

Geospatial-oriented

When information can be linked to a geographical location by means of coordinates or place names.

Geocoding required to convert addresses or city names into geographical coordinates.

FDI data from www.fdimarkets.com (FT)

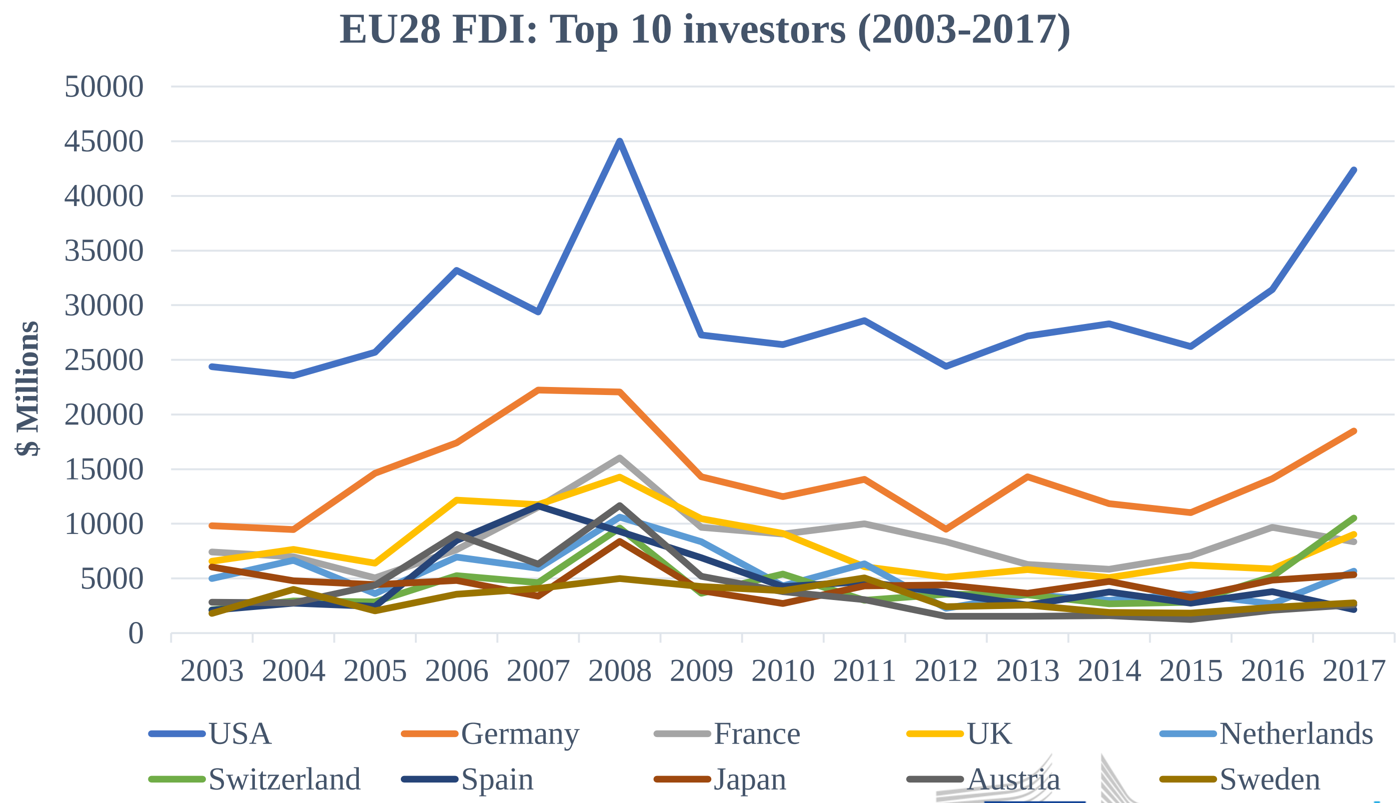
projectsYear	sourceCntr	sourceState	sourceCity	destCntr	destState	destCity	subsidiaryCompany	cluster	activity	subSector	typeFDI	market	motive	capitalEx	Jobs	signal
Feb-18	Czech Republic	Czech Republic	Humpolec	Romania	Romania	Pitesti	CTP Invest	Construction	Construction	Industrial	Expansion			$ 108.50 m	* 2,775	Feb 19 2018
Feb-18	Czech Republic	Czech Republic	Humpolec	Romania	Romania	Timisoara	CTP Invest	Transportation, Warel	Construction	Industrial	Expansion			$ 108.50 m	* 2,775	Mar 05 2018
Feb-18	Czech Republic	Czech Republic	Humpolec	Romania	Romania	Floresti	CTP Invest	Transportation, Warel	Construction	Industrial	New			$ 108.50 m	* 2,775	Mar 05 2018
Feb-18	United States	California	Long Beach (CA)	Cuba	Ciudad de La Habana	Havana City	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New	Domestic	Proximity to markets or customers	$ 0.90 m	* 13	Mar 23 2018
Feb-18	United States	California	Long Beach (CA)	Cuba	Camagüey	Camagüey	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New			$ 0.90 m	* 13	Mar 06 2018
Feb-18	United States	California	Long Beach (CA)	Cuba	Cienfuegos	Cienfuegos	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New			$ 0.90 m	* 13	Mar 06 2018
Feb-18	United States	California	Long Beach (CA)	Cuba	Matanzas	Matanzas	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New			$ 0.90 m	* 13	Mar 06 2018
Feb-18	United States	California	Long Beach (CA)	Cuba	Santiago de Cuba	Santiago de Cuba	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New			$ 0.90 m	* 13	Mar 06 2018
Feb-18	Cuba	California	Long Beach (CA)		Ciudad de La Habana	Havana City	Cuba Travel Services (CTS)	Tourism	Sales, Marketing & Support	Travel	New			$ 0.90 m	* 13	Mar 06 2018
Feb-18	United States	New York	NYC (NY)	Czech Republic	Czech Republic	Brno	Cushman & Wakefield	Professional Services	Business Services	Real estate	New	Domestic	Domestic Market Growth Potential	$ 35.90 m	* 20	Feb 16 2018

Multidimension dataset on FDI:

Spatial: Source – Destination (country/region/city).
Temporal: Monthly for 2003-2018.
Thematic: Sector, activity, type, market, motive…
Capital expenditure and **Jobs created.**

EU28 investors: most common value (frequency)

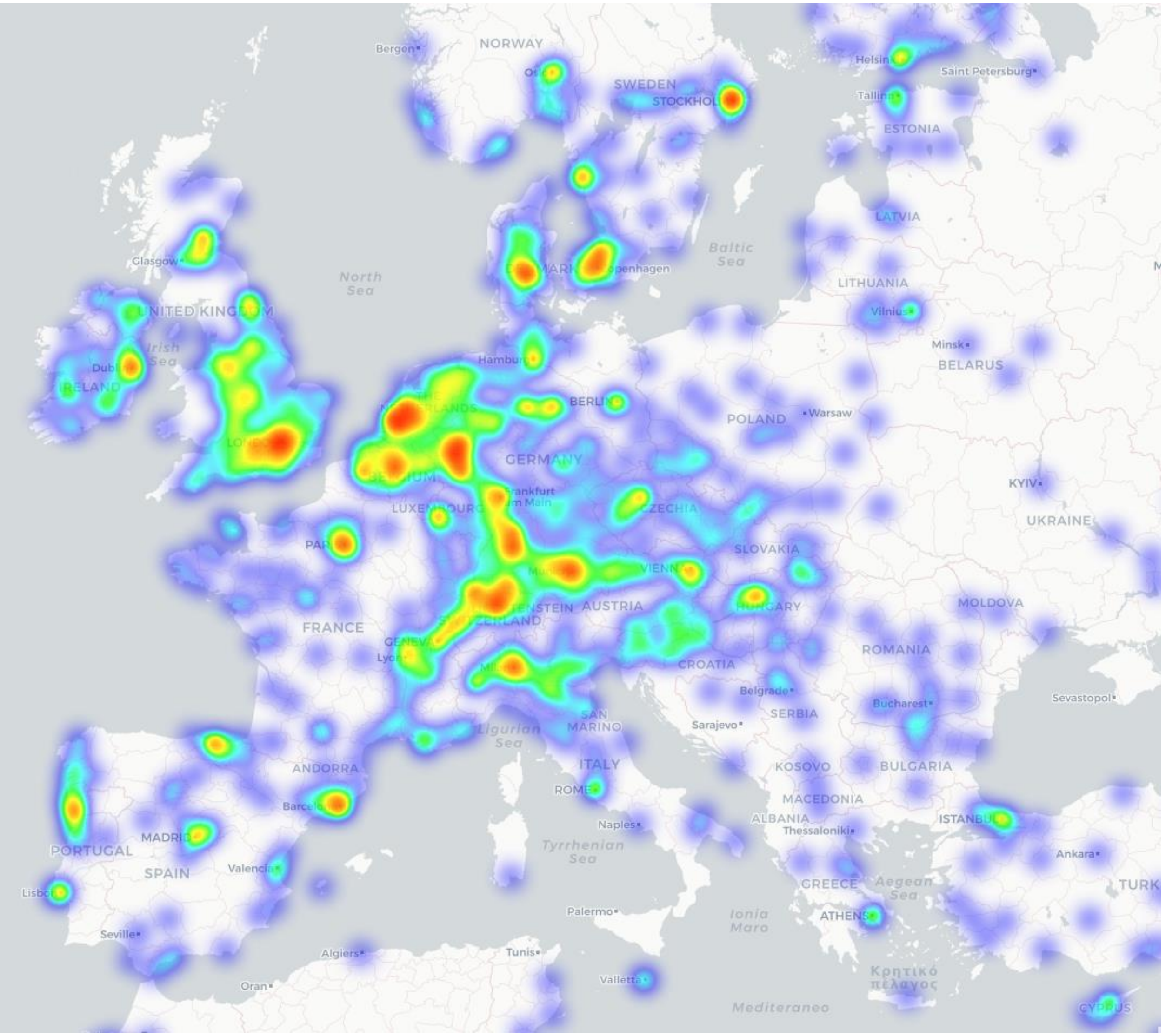
Period	Source Country	Sub Sector	Market	Motive
2003-2018	United States (14576)	Software publishers, except video games (5279)	Regional (8028)	Proximity to markets or customers (1357)



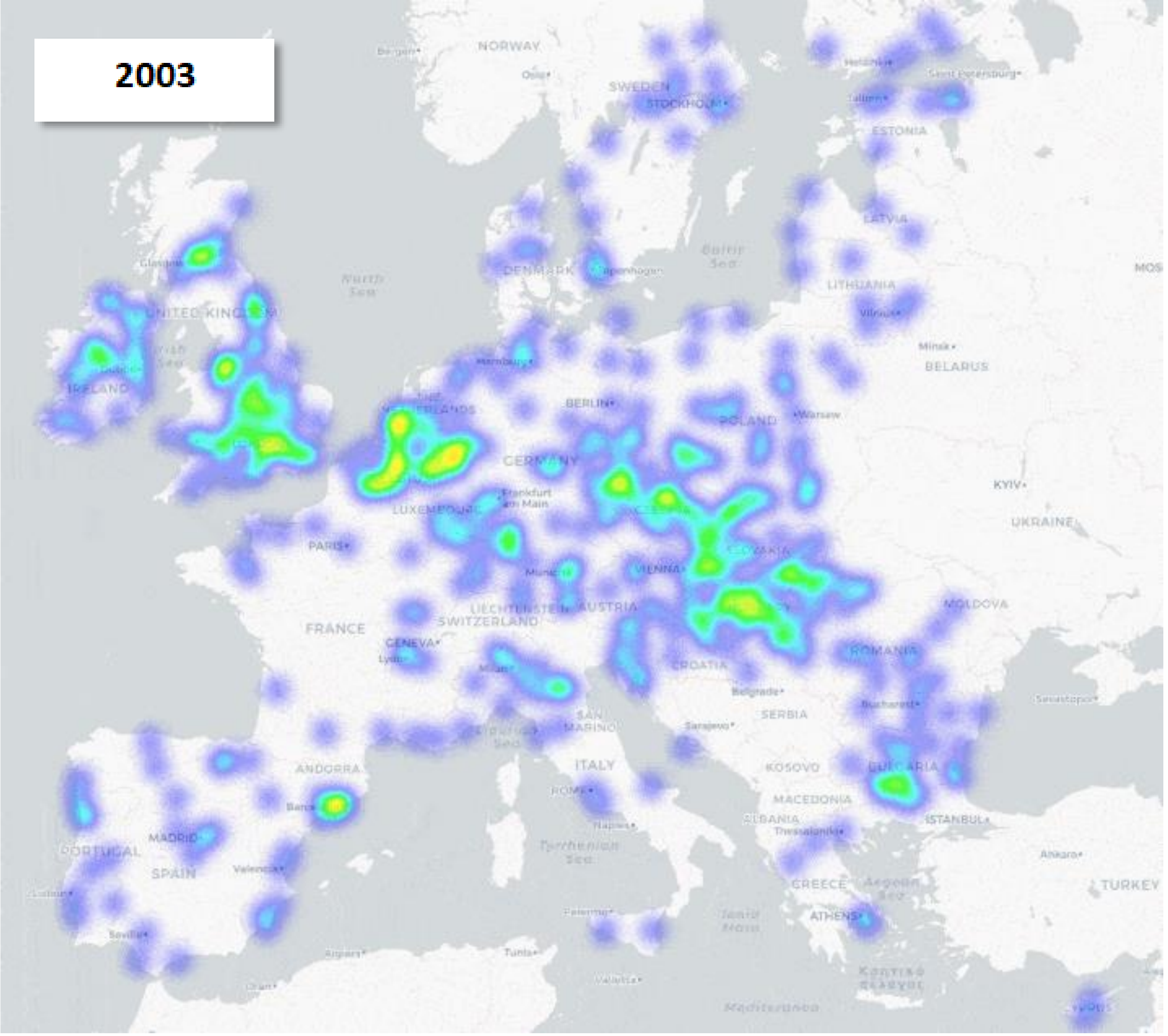
European Commission

Mapping hot spots of FDI origin and destination

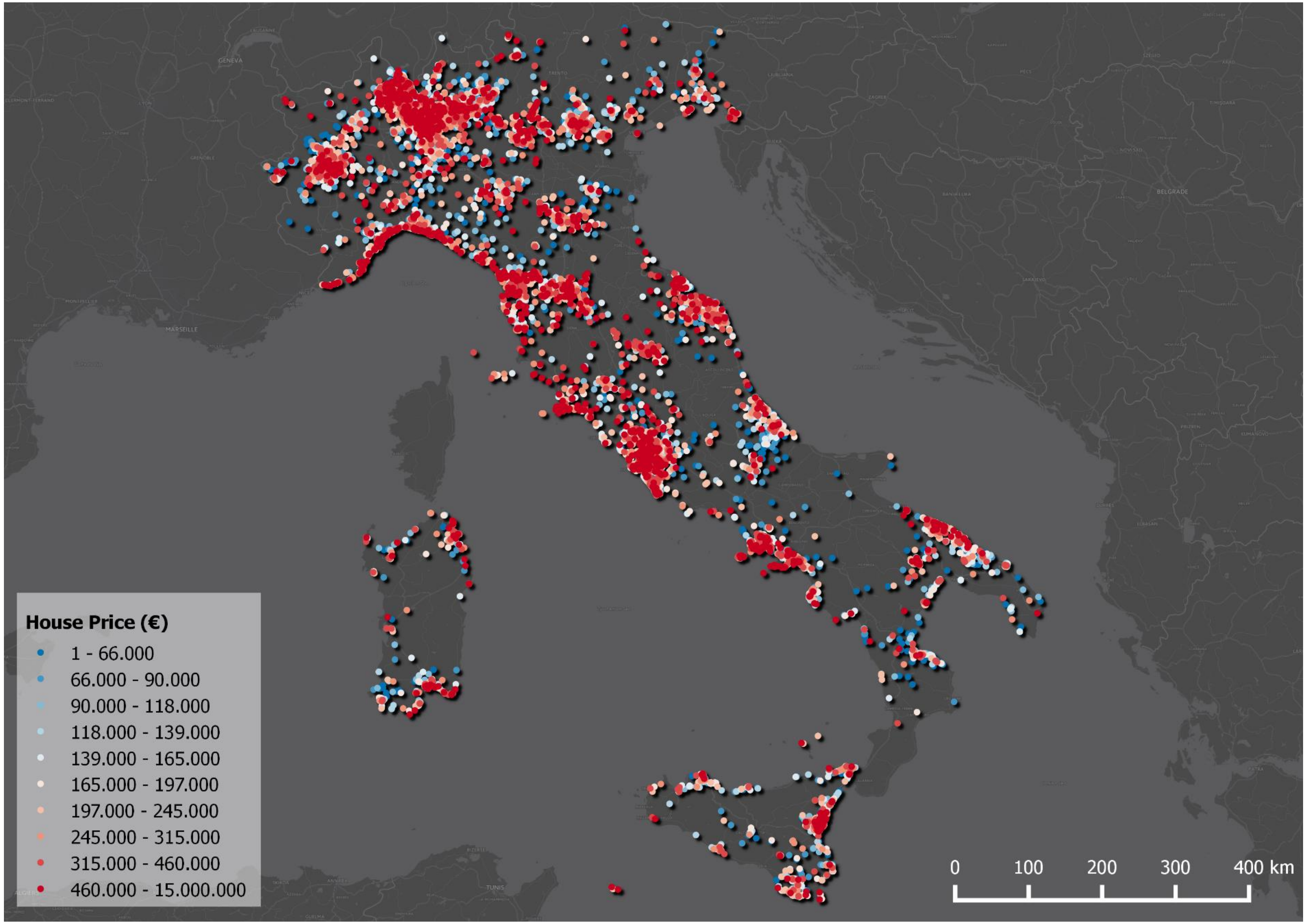
Source (2003-2018)



Destinations



Housing (ask) prices from Remax (Italy)



Features (Total 34):

lat - Latitude
lon - Longitude
keys - ID
price - Selling Price (€)
totalRooms - Total Rooms
bedrooms - Bed Rooms
bathrooms - Bathr
totalsqm - Total Square Meters (m2)
lotsize - Lot Size (m2)
year - Construction Year
builtArea - Building Area (m2)
parkingspaces - Number of parking spaces
floors - Number of floors
floorlevel - Floor of the house
toiletRooms - Number of toilets
energyClass - Energy class ***
energyEff - Energy Efficiency (kWh/m2 per year) ***

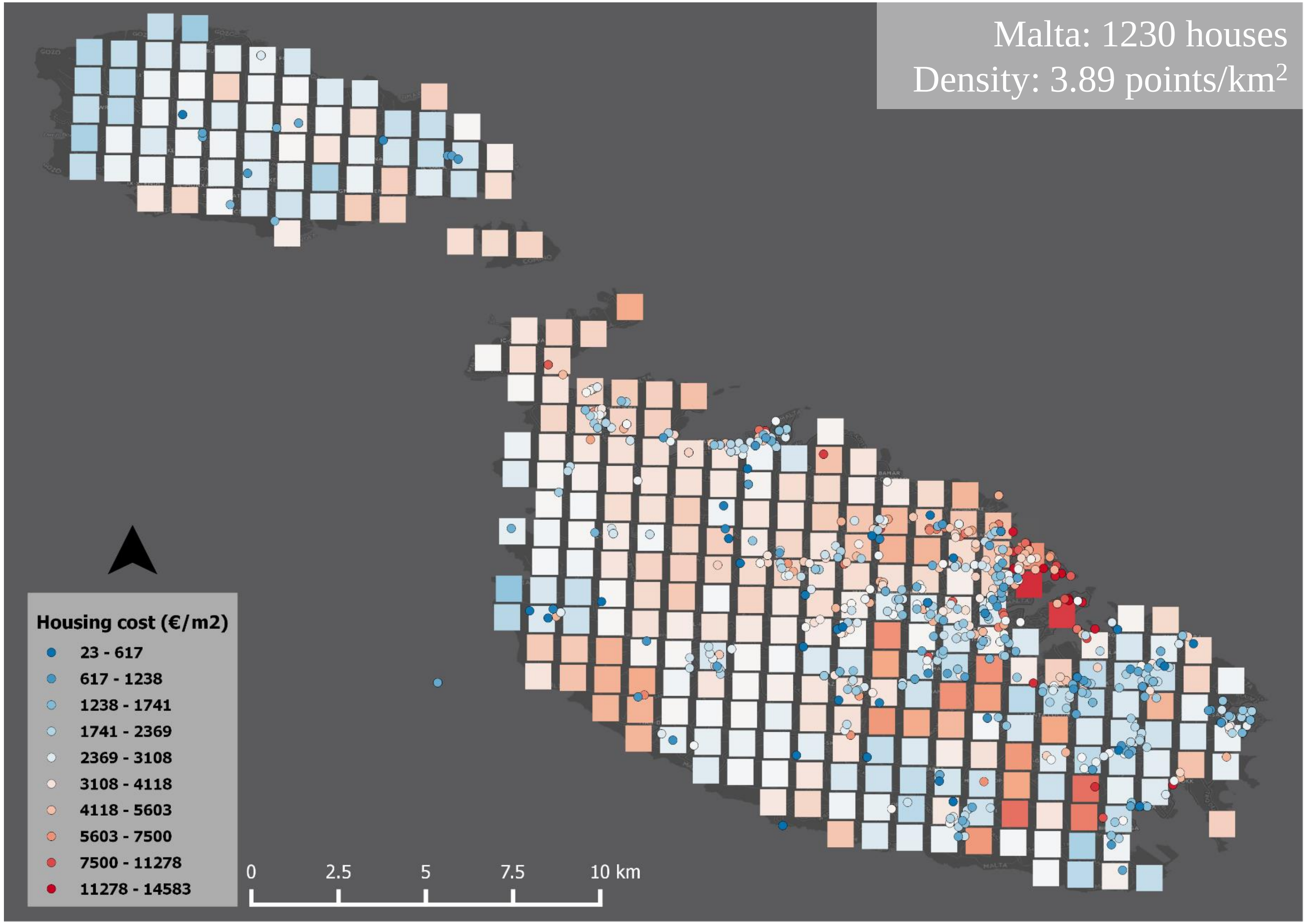
Features (Total 17 - yes or no):

garage - Garage (yes/no)
pool - Swimming Pool (yes/no)
renovated - Renovated (yes/no)
fireplace - Fireplace (yes/no)
terrace - Terrace (yes/no)
balcony - Balcony (yes/no)
garden - Garden (yes/no)
liftelev - Lift or Elevator (yes/no)
parking - Parking places (yes/no)
heating - Heating System (yes/no)
solar - Solar panels (yes/no)
oil - Oil heating (yes/no)
ac - Air Conditioner (yes/no)
sewer - Connected to sewer (yes/no)
pool - Swimming Pool(yes/no)
security - Alarm or security system (yes/no)
kitchen - Kitchen (yes/no)

European Commission

Housing (ask) prices from Remax (Malta)

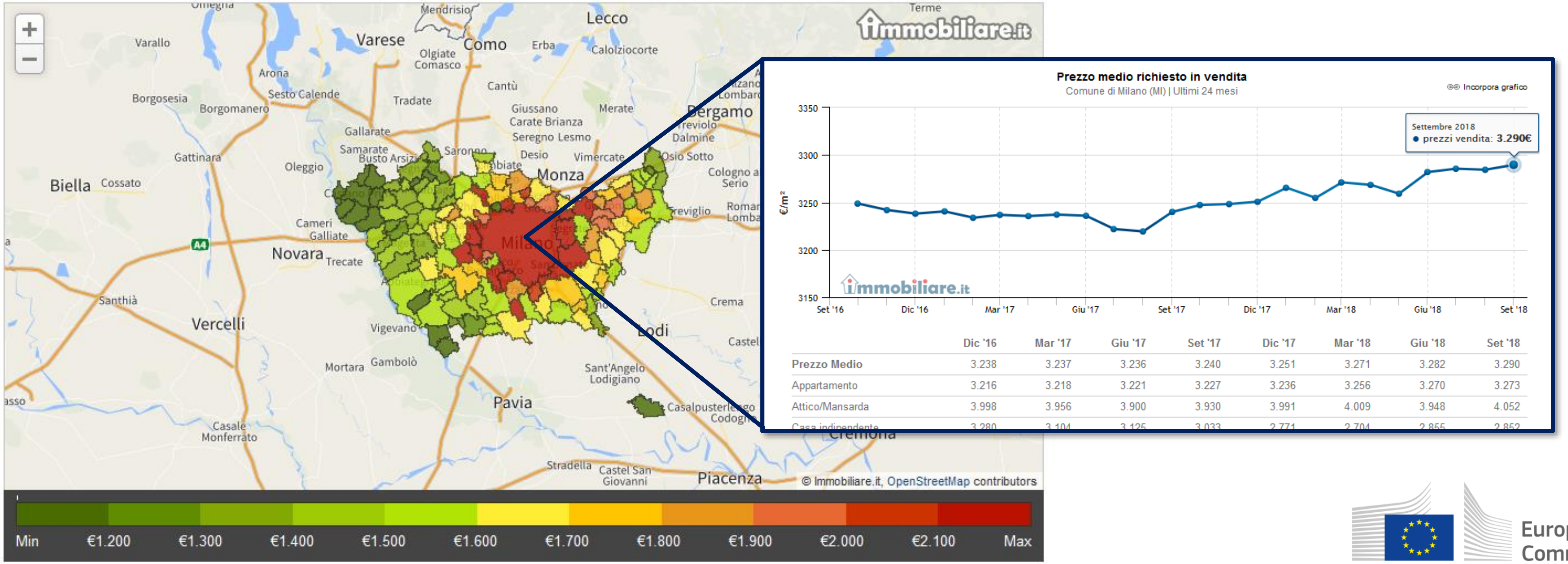
- The collection of refined housing data combined with local and neighbourhood characteristics, allows the development of advanced regression models.
- These can be use to analyse the most important factores driving house prices, make predictions and create spatial cost grids.
- On-going: selection of European cities for refined case-studies.



Housing (ask) prices: Milan time-series (Sep 2016–Aug 2018)

Data source: immobiliare.it

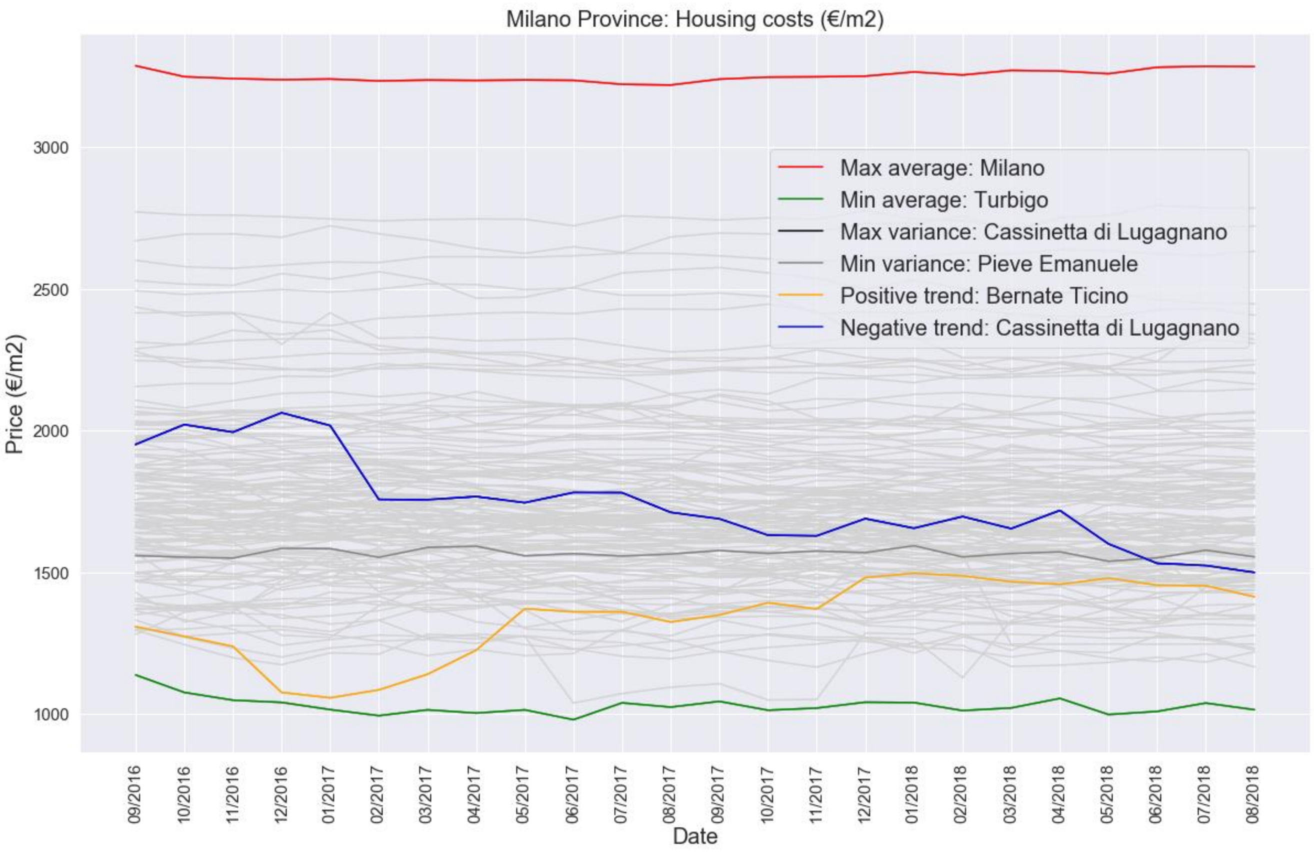
- Italian website that collects selling/rent housing data from several real-estate agencies.
- Wide thematics, statistical and spatial coverage.
- Collects average monthly sell prices for the last 2 years.

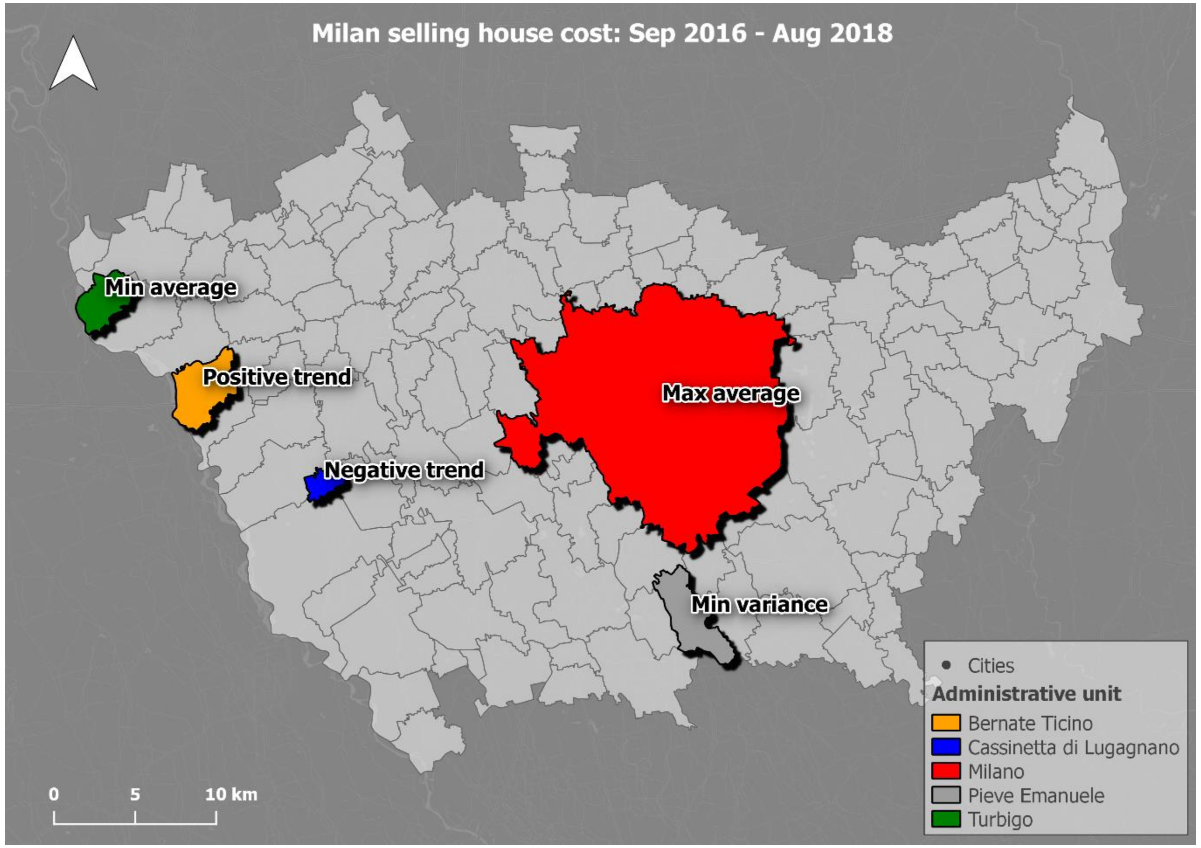


	Dic '16	Mar '17	Giu '17	Set '17	Dic '17	Mar '18	Giu '18	Set '18
Prezzo Medio	3.238	3.237	3.236	3.240	3.251	3.271	3.282	3.290
Appartamento	3.216	3.218	3.221	3.227	3.236	3.256	3.270	3.273
Attico/Mansarda	3.998	3.956	3.900	3.930	3.991	4.009	3.948	4.052
Casa indipendente	3.280	3.104	3.125	3.033	2.771	2.704	2.855	2.852

European Commission

Housing Cost: Milan time-series (Sep 2016 – Aug 2018)





European Commission

Towards spatiotemporal population

The ENACT project

- JRC Exploratory project (2016-17 + maintenance in 2018)
- To produce multitemporal population grids taking into account daily and seasonal population variations
- 1 Km^2 resolution, whole EU28, consistent and validated methodology

Essential to assess impacts and prepare strategies in various domains

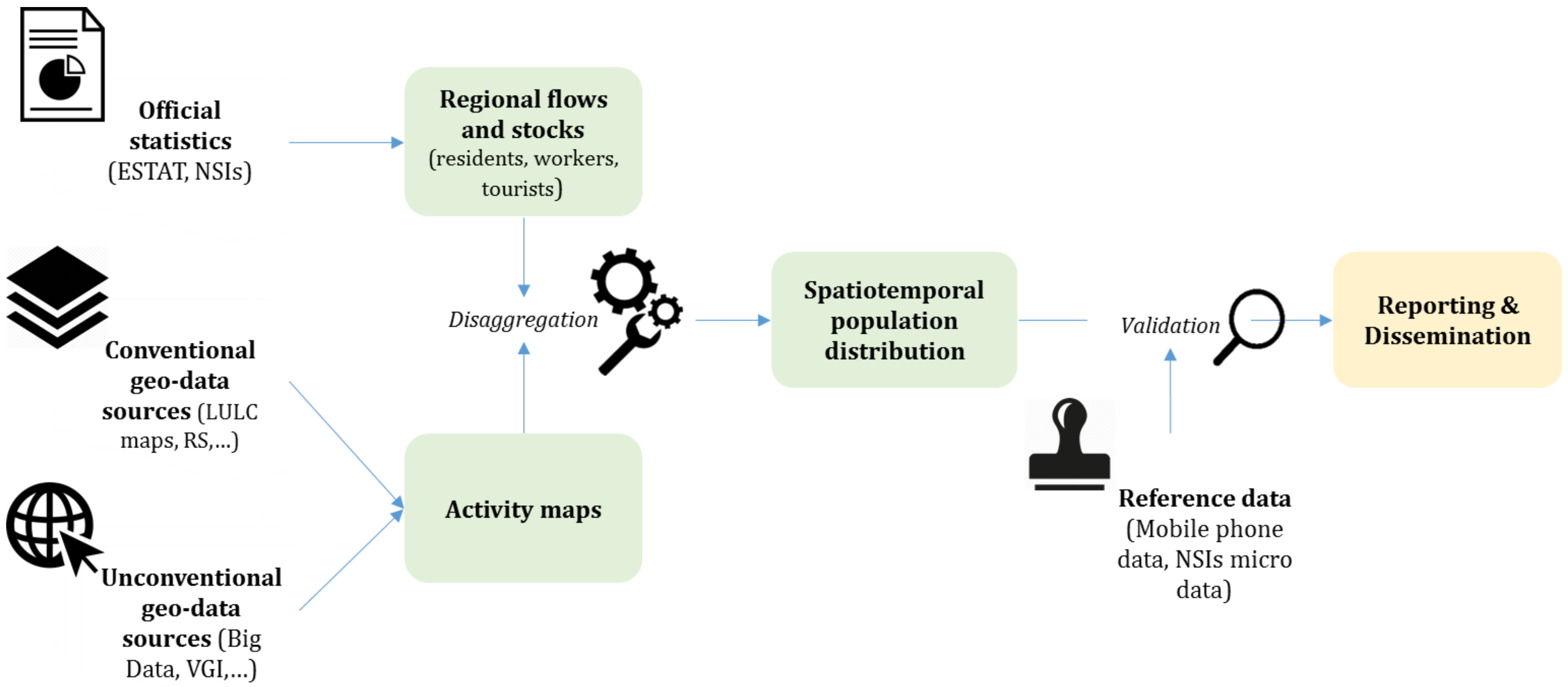
- Urban and regional planning, Disaster risk and emergency management, Exposure to pollutants, Epidemiology, Geomarketing…

Interesting example of how conventional and non-conventional data can be combined to generate a product with significant value added.

European Commission

Towards spatiotemporal population

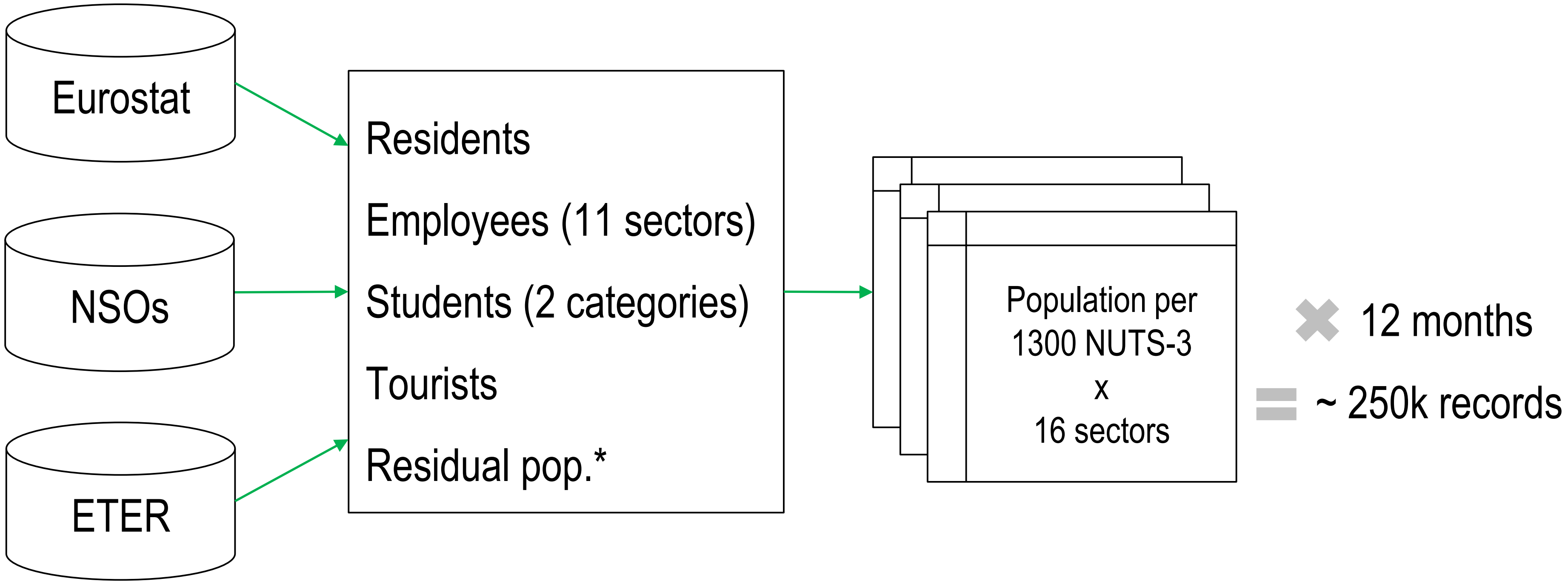
Workflow



European Commission

Regional population flows and stocks

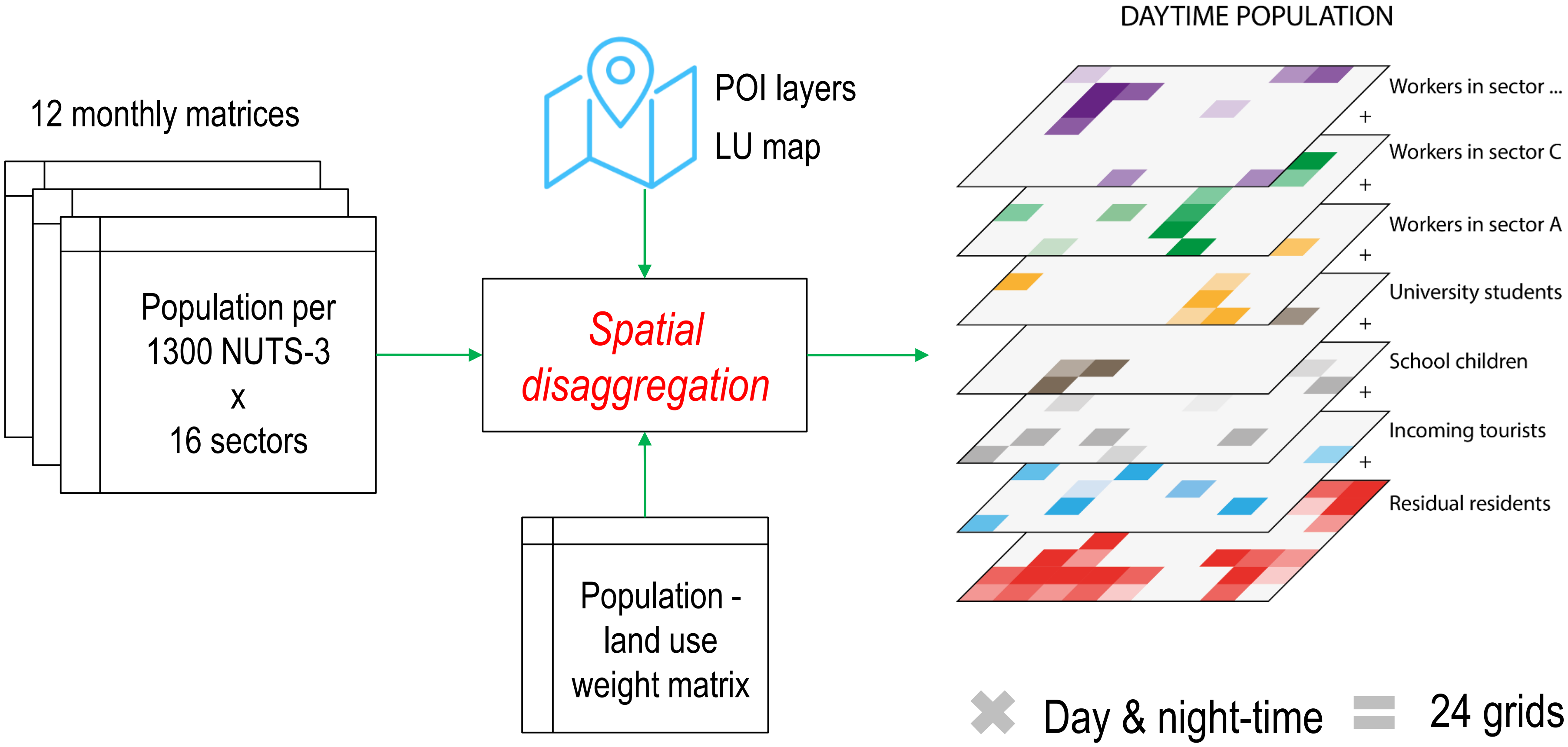
Estimation of flows and stocks of 16 population subgroups per NUTS-3 regions.



*population not working nor studying = retired + children + unemployed + inactive working age

Population disaggregation

Creation of population grids by disaggregating regional population stocks to grid level, using location of activities as spatial proxies.



Mapping tourism

Table 2. No. of establishments and no. rooms per data source for EU-28.

	No. of establishments	No. of rooms	Nr. bed-places
Booking.com	532,346	7,528,249	n.a.
TripAdvisor	310,958	9,818,732	n.a.
Combined	716,103	13,218,804	n.a.
Eurostat	597,358	n.a.	30,850,722

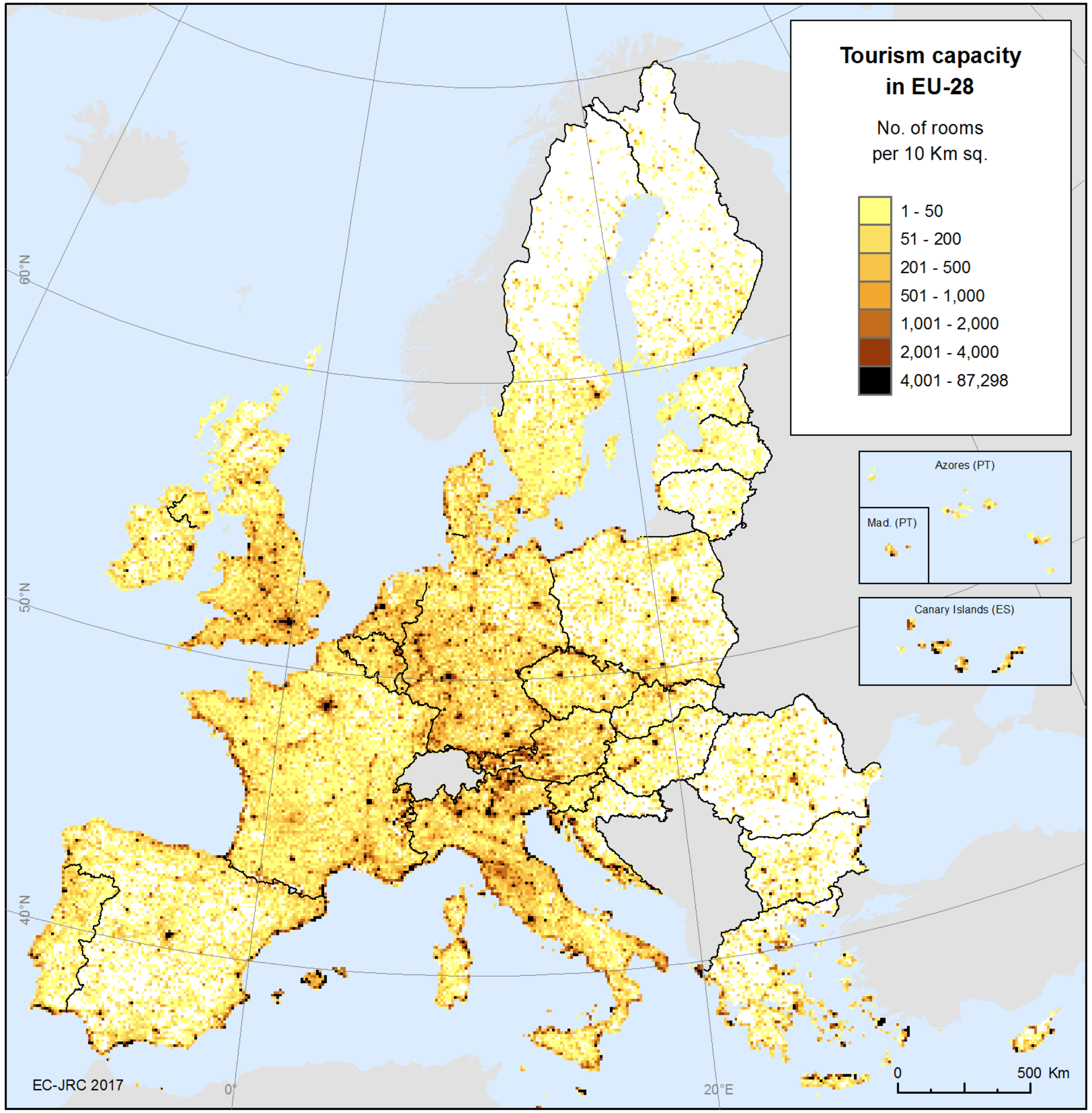
Notes:

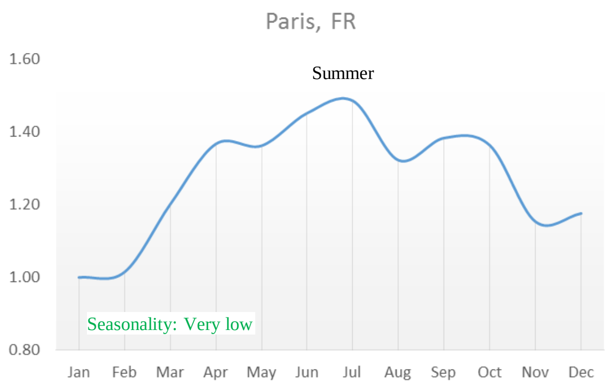
n.a. = not available

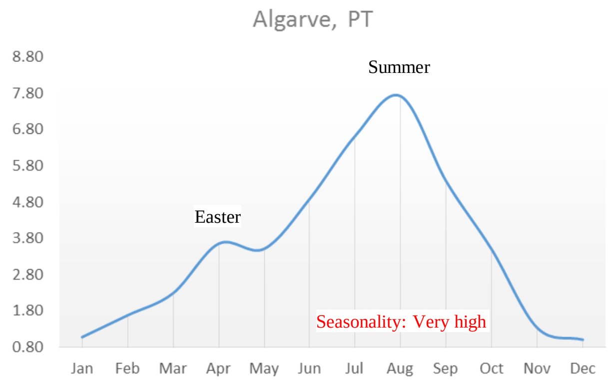
1) All values refer to the territory of EU-28, excluding Atlantic islands of Portugal and Spain and French overseas territories.

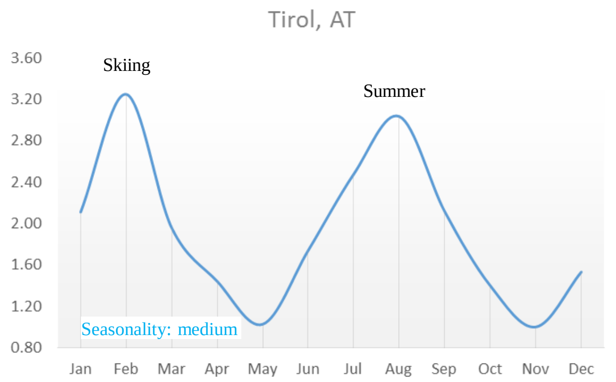
2) Figures from Eurostat refer to the year 2016, except for Ireland and Portugal (2015).

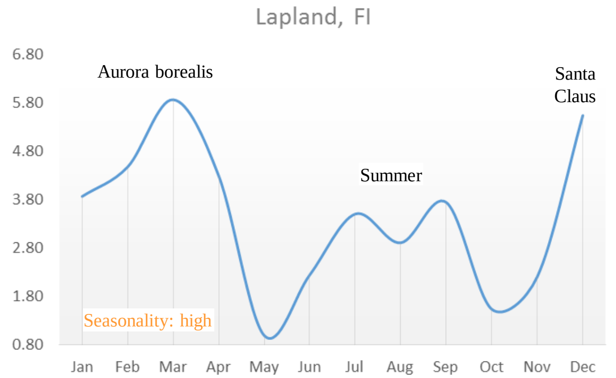
3) Figures regarding Booking.com and TripAdvisor as of February 2017 and August 2017, respectively.

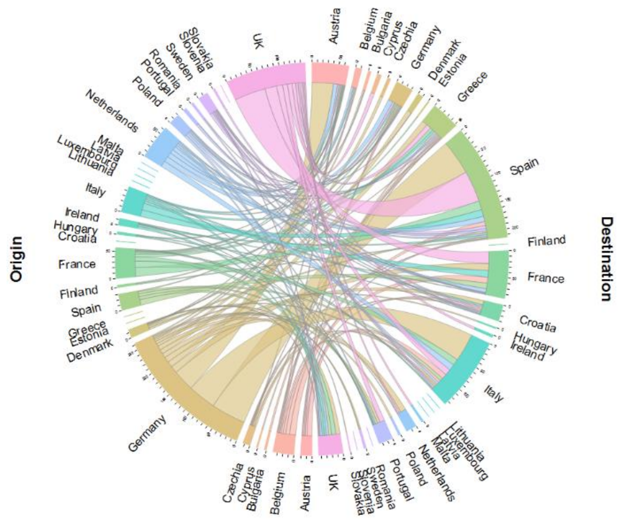


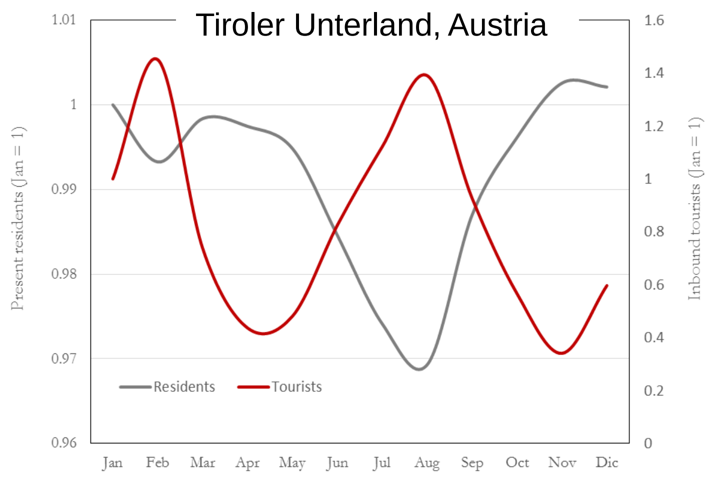


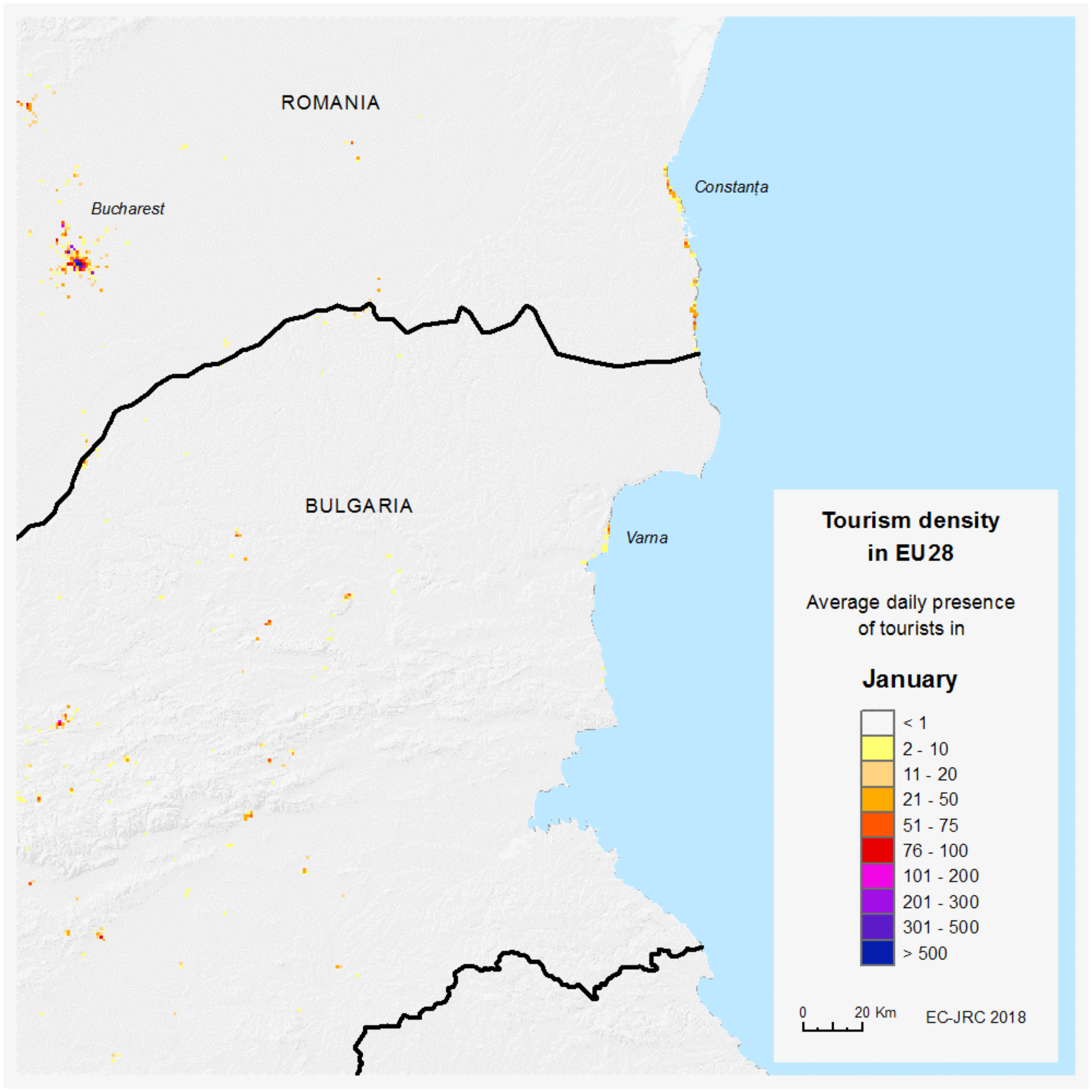




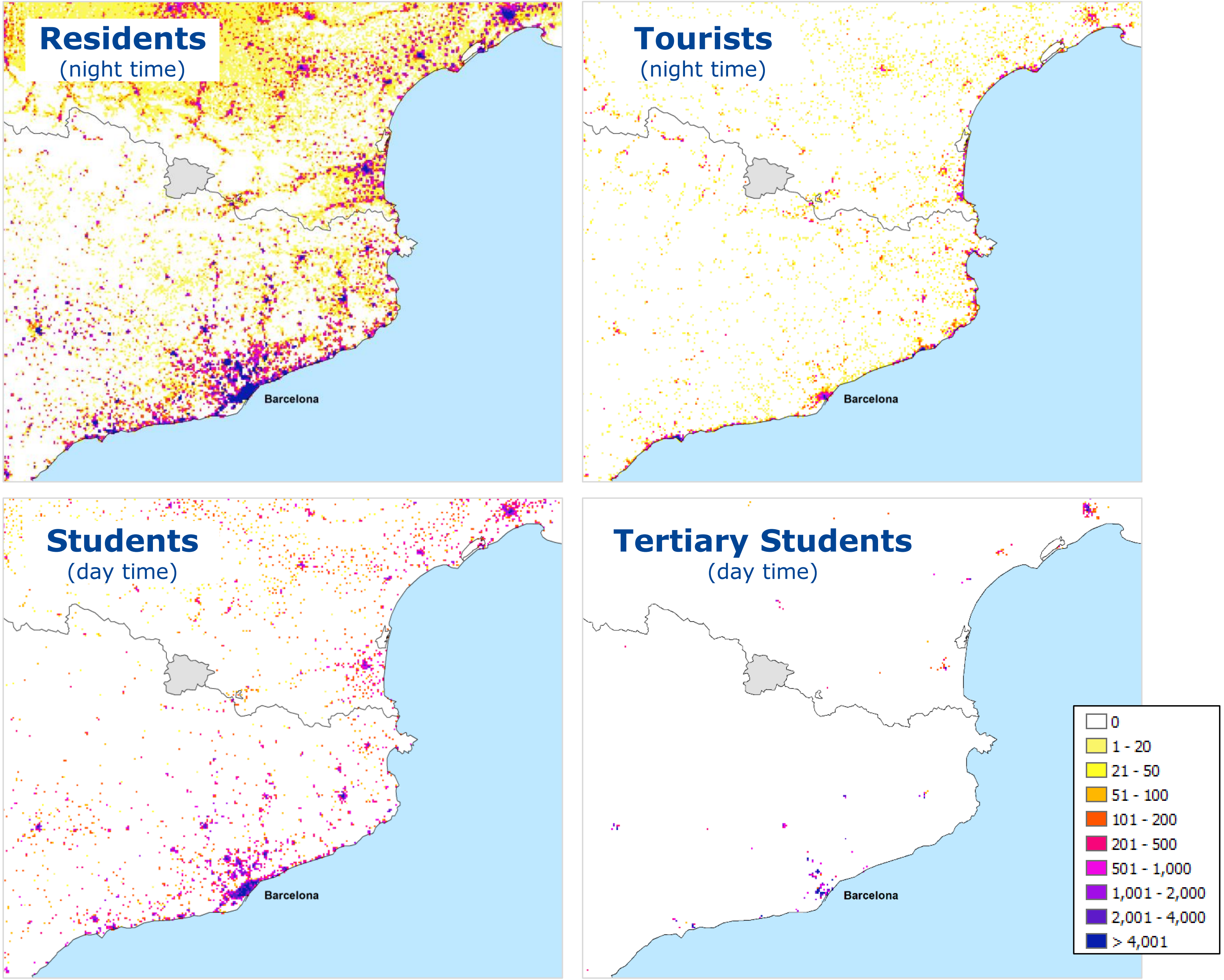






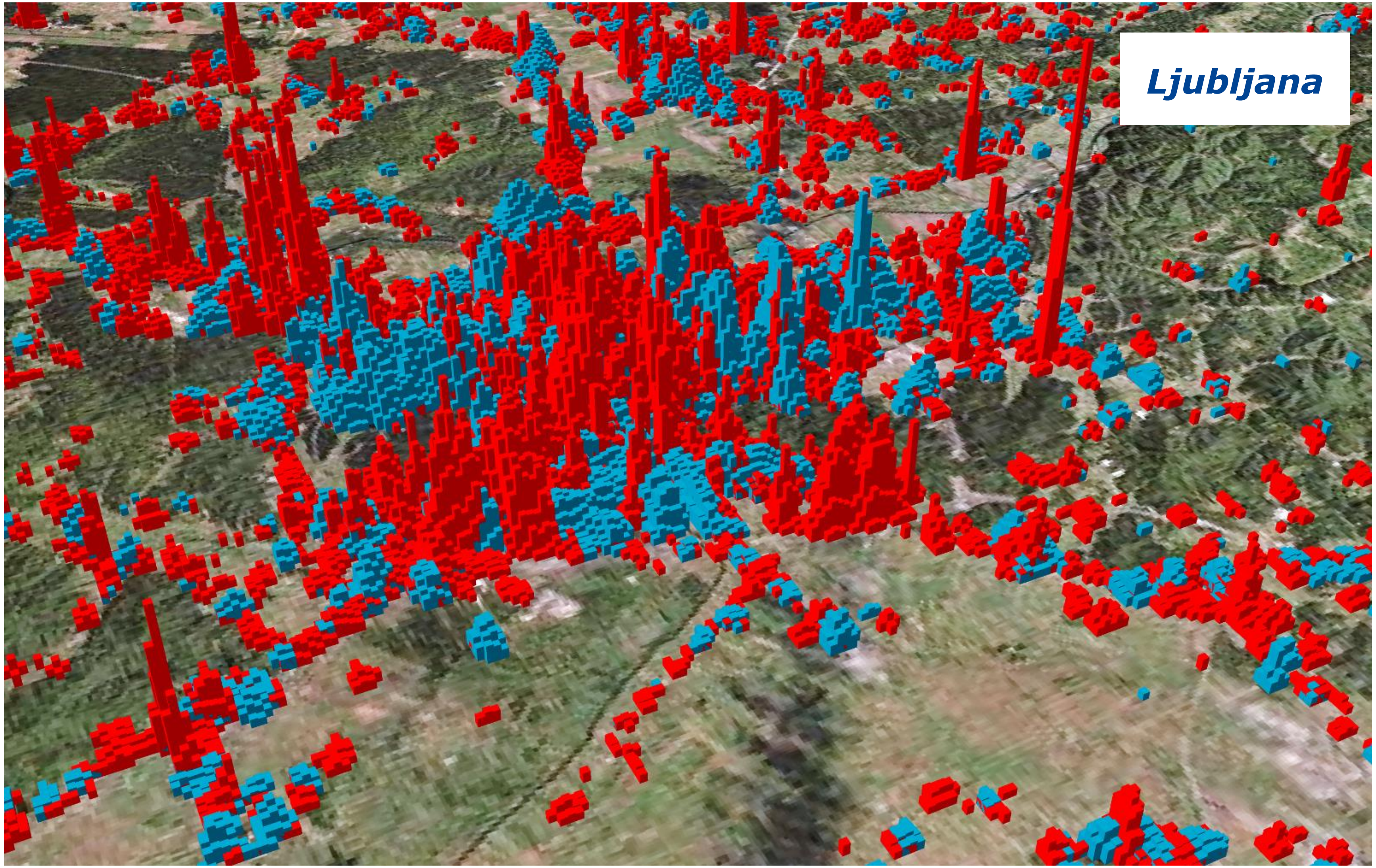


European Commission



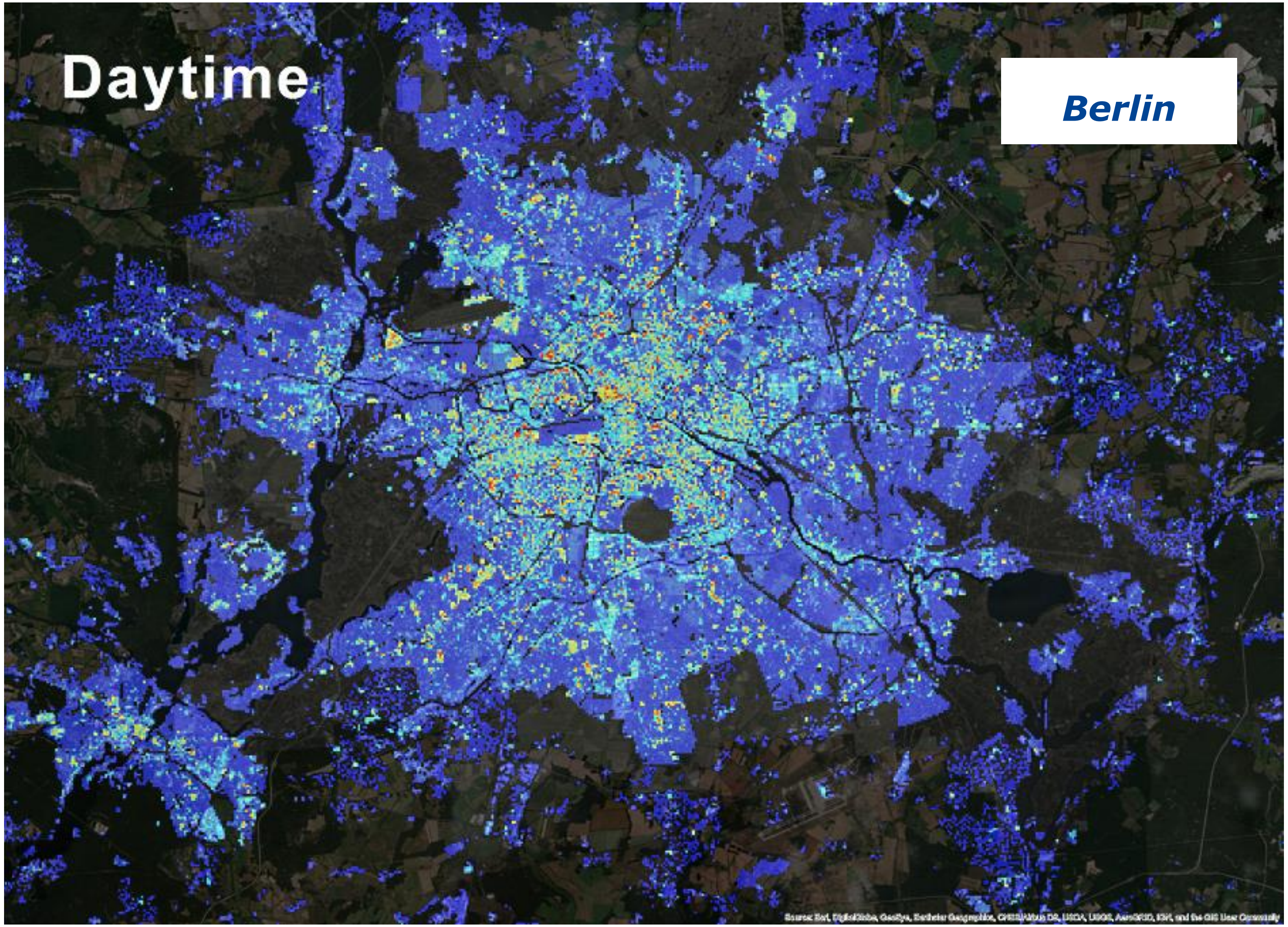
European Commission

Day and night-time population



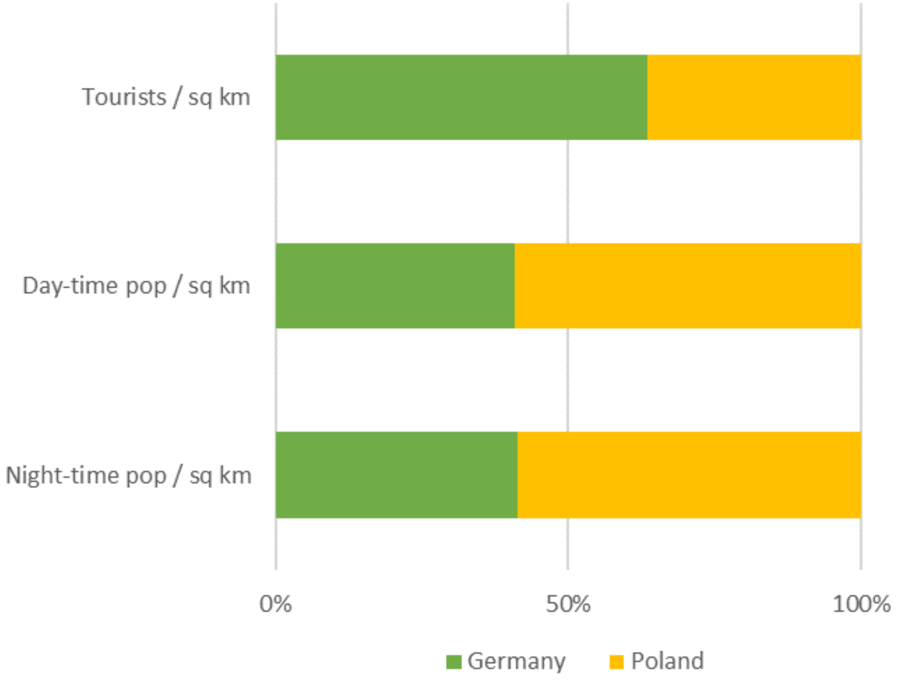
Night-time population density   Day-time population density

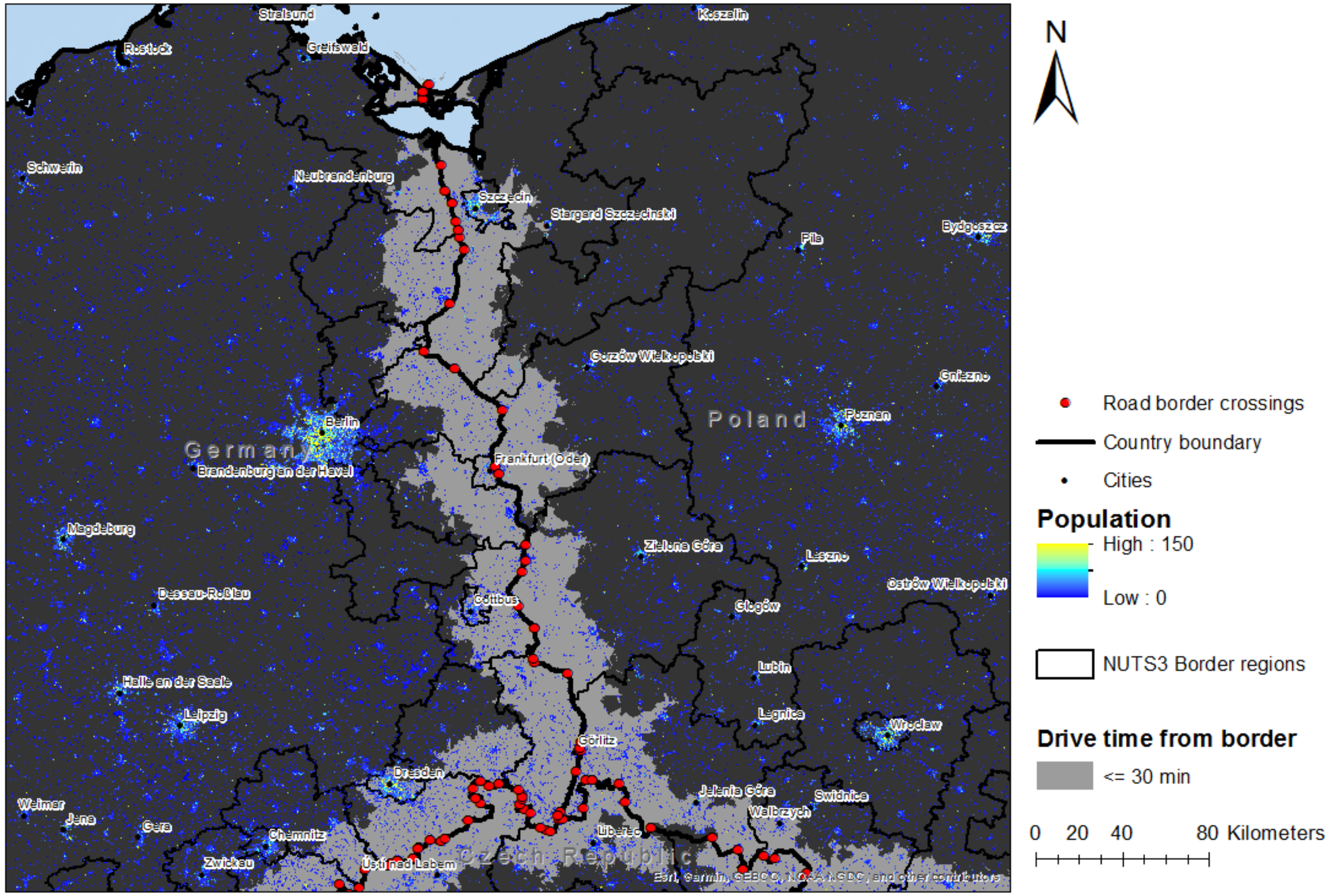
Day and night-time population



ENACT – use case

Characterization of functional cross-border areas





ENACT - way forward

On-going work

- Fine tuning model parameters
- Updating statistical data to year 2016
- Dissemination and applications

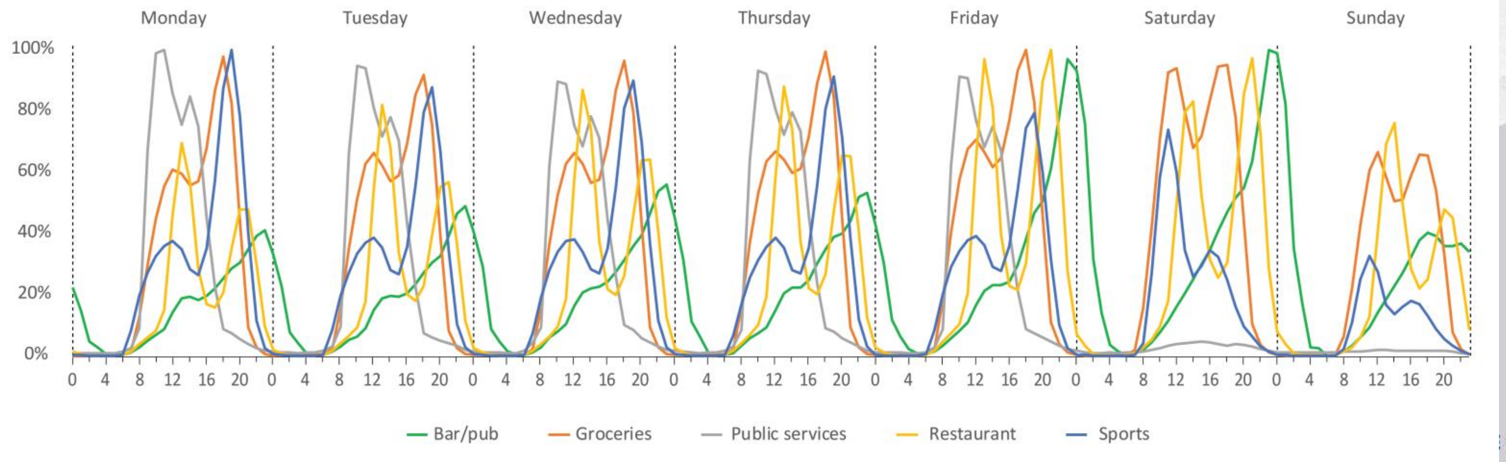
Way forward

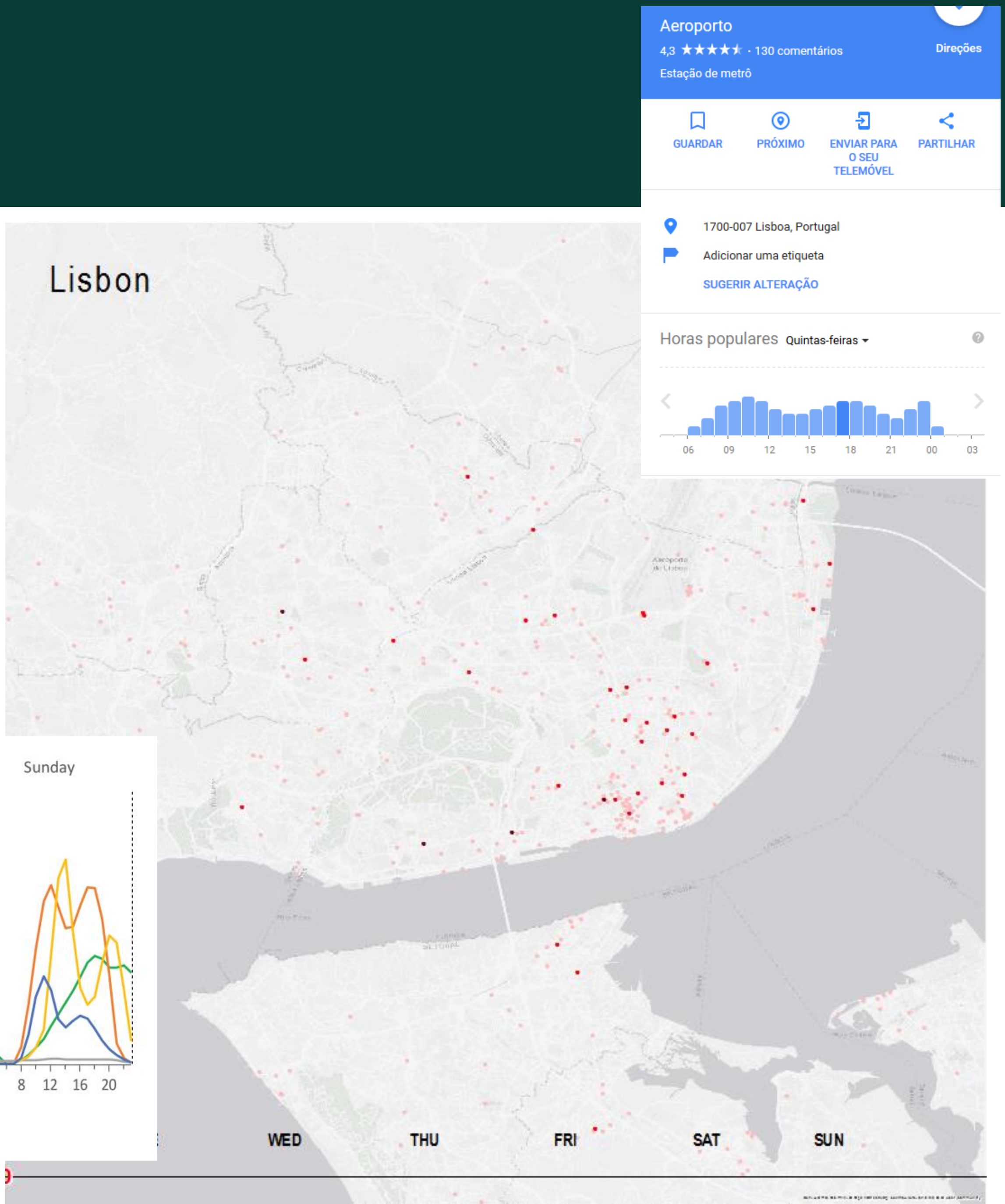
- Consolidate and automate aspects of the methodology
- Increase temporal resolution (from day/night to 24/7)

ENACT - way forward

New set of POI data from Google Maps enriched with Popular Hours data.

- Fine spatial resolution and 24/7 temporal detail.
- Multiple activity types.





Discussion and conclusions

- Emerging sources of geospatial data are promising inputs to **complement** (not replace) traditional/official sources.
- **Sustainability issues**
  - Many sources **may not be sustainable** in the long-run. Business/profit oriented ICTs.
  - Conditions that allowed the generation of the data (e.g. technology, market conditions, legal frameworks) may cease to exist or evolve in different directions.
  - Data access constraints (technological, legal, …).

European Commission

Discussion and conclusions

- **Quality issues**
  - Unlike NSOs, no mission to produce complete, consistent and frequent statistics.
  - Difficult to assess (no benchmarks). Use quality check by sampling methods, or systematic but indirect approaches.
  - Suitability of a data source depends on application.
  - Completeness, accuracy, semantic and ontological differences across different sources challenging to reconcile.

Thank you

Any questions?

You can find me at filipe.batista@ec.europa.eu