Measuring land use changes by (machine) learning from historical maps

The emergence, growth, and stagnation of cities: France c. 1760-2020

Pierre-Philippe Combes
Sciences Po, Paris

Gilles Duranton
University of Pennsylvania

Laurent Gobillon
Paris School of Economics

Clément Gorin
University of Toronto

Introduction: Objectives

- To track land use and urbanisation in France over 250 years.
- Part of a rising interest for historical data in urban economics/economic geography (Hanlon and Heblich, 2022; Combes, Gobillon, and Zylberberg, 2022).
- Series of projects:
  - Urbanisation in France over 1760-2020.
  - Land use change in France 1860-2020.
  - Structural change and urbanisation.

Introduction: Contributions

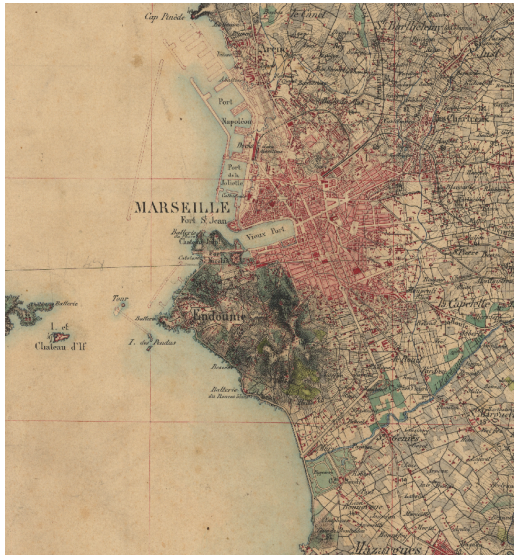
- Methods: Develop a new methodology to extract information from old maps using a combination of image processing and machine learning tools.
- Data production: Unique urbanisation and land use gridded data for France and 4m$\times$4m pixels over 1760-2020.
- Note: Economists are not interested in precisely describing what occurred in a given place (or 2 or 5 or 10), but on what happened on average over the whole of an economy (France here).

Main sources: Four series of French historical maps since 1760

Cassini, '1760'

First ever using triangulation for coverage of an entire country

1/86,400



Military, '1860'

Military maps

Many land use categories

1/40,000

Scan50, '1960'

From aerial photographs

More symbols, fewer colours

1/50,000

2020

Actual geocoded information for most land uses from various sources

Machine-learning strategy

- 1760, built-up only:
  - Cassini project's encoding of main cities and symbols (churches, castles, mills),
  - Also use of 18^{th} c. buildings that still exist nowadays (CEREMA),
  - Random forest to predict other built-up areas (small towns, villages).
- 1860, 7 land uses (aggregation of the 57 represented on maps):
  - A combination of many random forests strategies,
  - With some pre- and post-processing.
- 1960: built-up only.
  - More complicated due to much less coloured information.
  - Use of convolutional neural networks.
- 2020: Actual geo-coded data for all land uses.

1860 land use extraction

- Great job, not done by us, in precisely scanning and geocoding the paper maps.
- Main issue: georeferenced images (geotiff) only: Each pixel's coordinates and *rgb* values (colour) but no label corresponding to its land use.
- Other issues:
  - 33.86 billion of 4m$\times$4m pixels to be coded.
  - Different shading / damaged parts across different maps.
  - Overlaying (names, level contours), identical colours for different objects.

$\Rightarrow$ Need to combine many image processing and machine learning techniques to

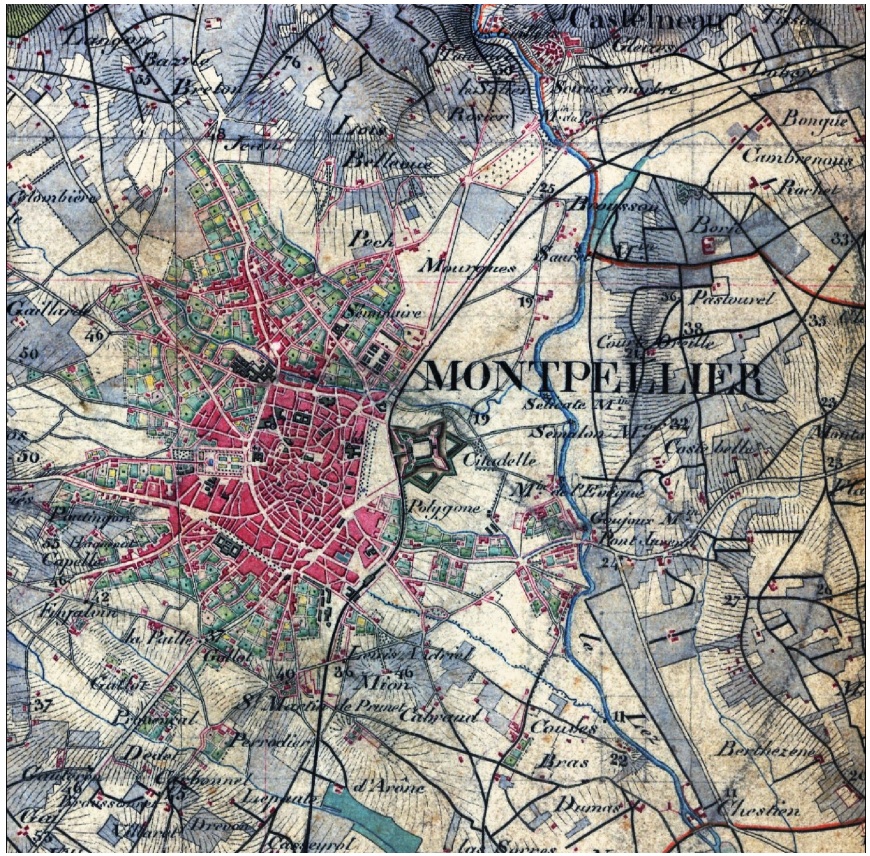
Outline of the machine-learning strategy for 1860

Multi-step methodology mostly based on 'Random Forests':

1. Image pre-processing to homogenize colours and augment contrast.
2. Separation of built-up vs all other land uses at the 4m$\times$4m pixel level:
   - Random forest 1: Reddish pixels vs. all other land uses.
   - Post-processing: Remove walls (also isolated pixels; also fills small holes).
   - Random forest 2: Built-up vs others for reddish pixels.
3. Land use classification within not built-up pixels:
   - Clustering: Aggregation of neighbouring similar pixels into 'superpixels' using a Quickshift procedure.
   - Random forest 3: Superpixels classification in six land uses.

Visual results, built-up
Pre-processing, Montpellier

Original map



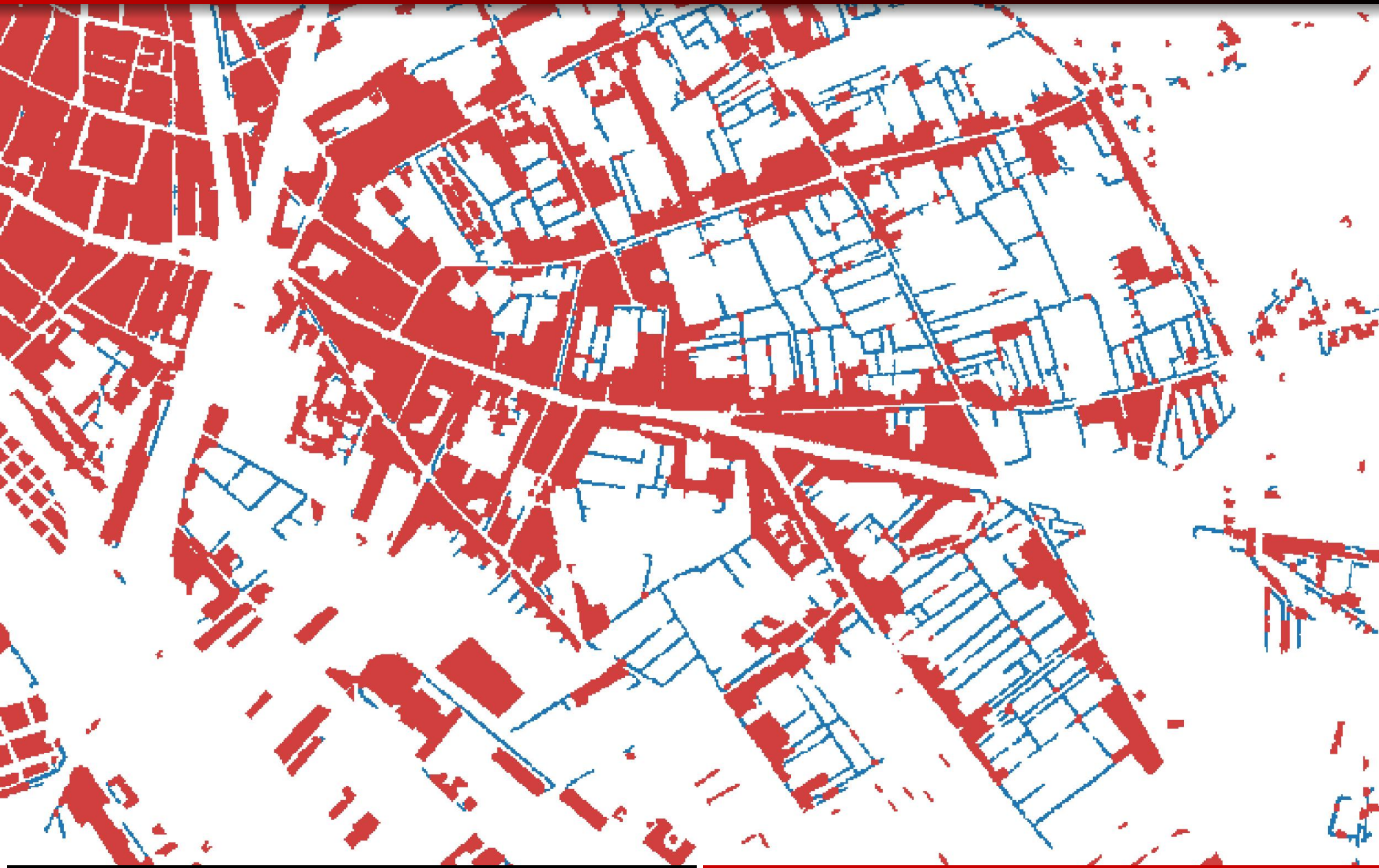
Pre-processed map

Reddish parts vs. the rest, l'Arbresle

Original map

Reddish pixels

Removal of walls within reddish pixels with post-processing



Built-up vs. reddish parts, l'Arbresle

Reddish pixels

Built-up vs other reddish pixels

Built-up vs. the rest, Marseille

Raw image

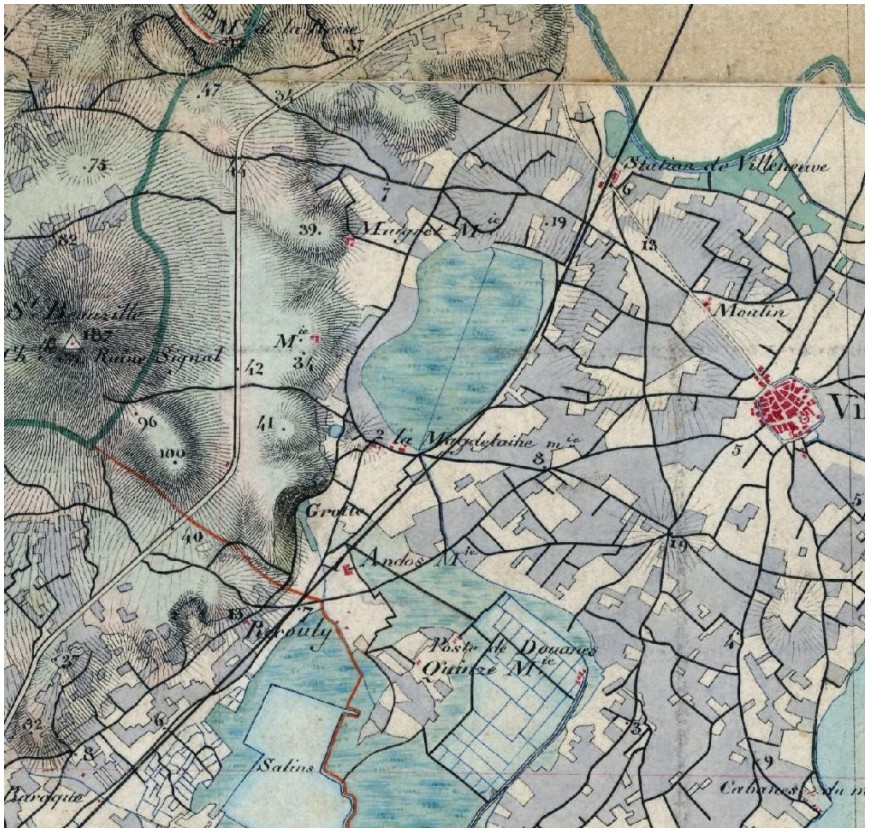
Built-up

Built-up vs. the rest, Lyon

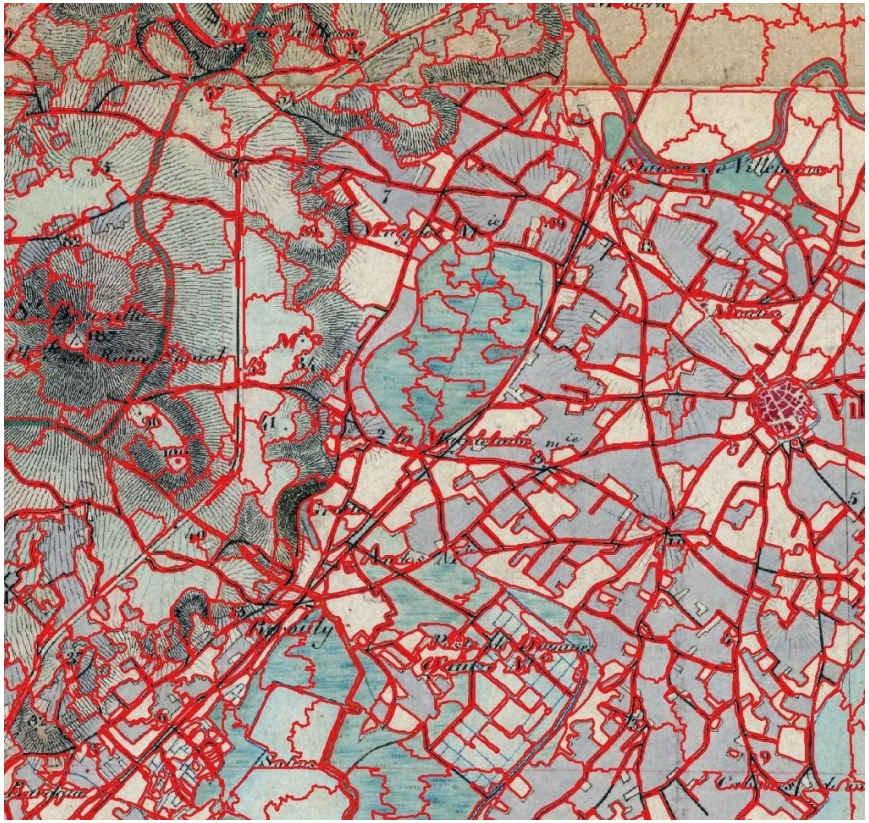
Raw image

Built-up

Superpixels from the Quickshift procedure

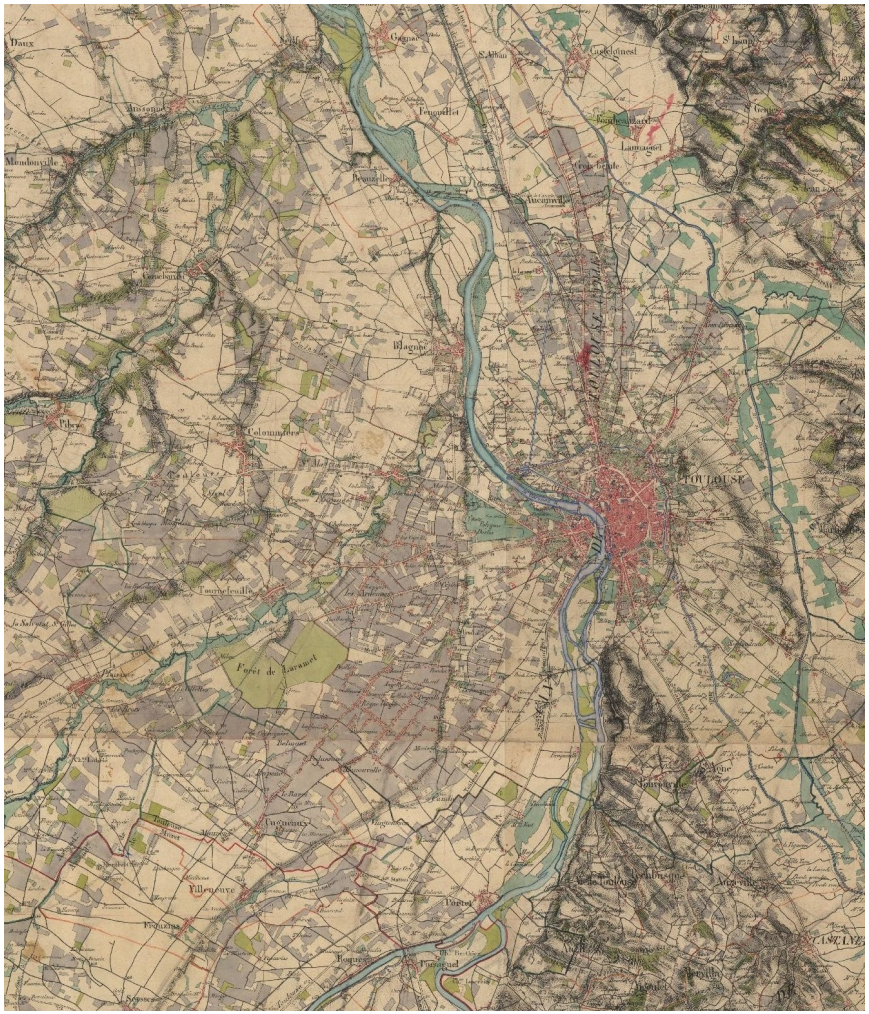


Original (pre-processed) map

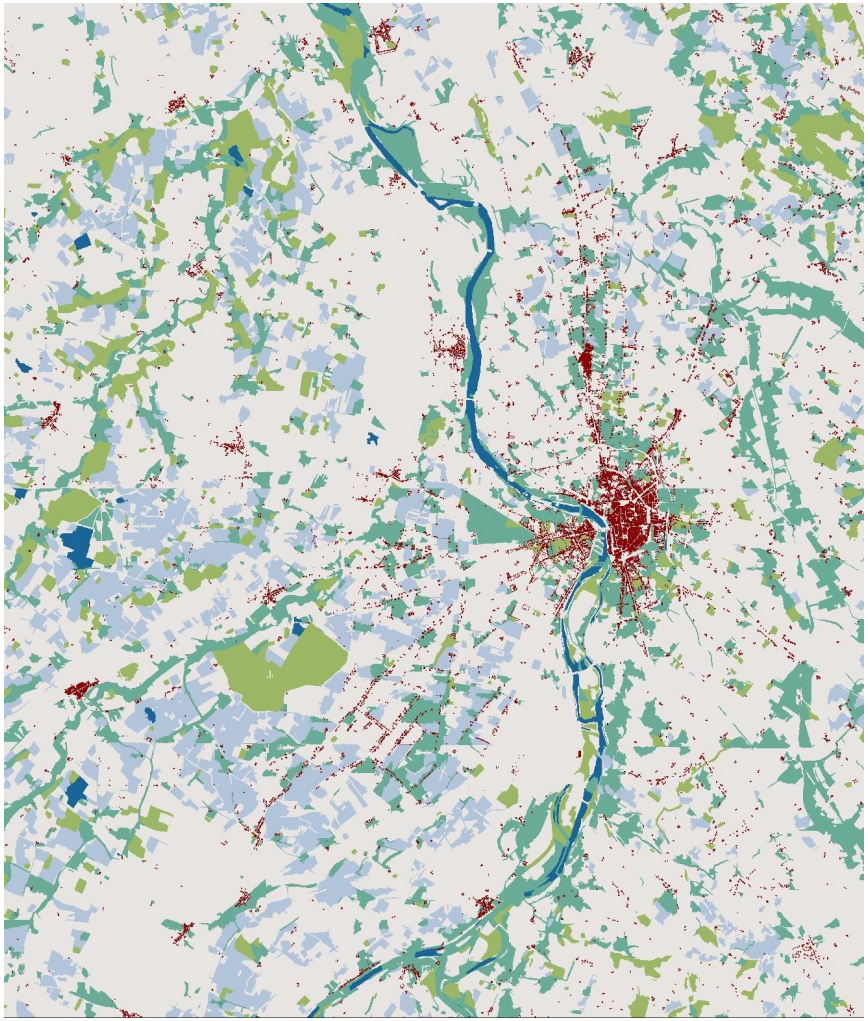


Superpixels from Quickshift

All land uses prediction, Toulouse



Raw image



Land uses

Prediction rates for all land-uses

- Prediction rates on all the pixels manually classified:

	Built-up	Crops	Meadows	Pastures	Specialised	Forests	Water
All pixels							
Recall	94.6%	95.5%	84.5%	90.3%	80.2%	93.1%	79.8%
Precision	85.8%	93.5%	87.9%	94.2%	92.2%	92.9%	74.0%
Without borders							
Recall	99.3%	98.9%	93.1%	95.5%	87.9%	97.4%	99.6%
Precision	96.1%	97.2%	96.7%	97.8%	98.1%	97.6%	93.9%

- Overall share of correctly predicted pixels among the 6.2 billions pixels manually classified: 92.2%, 97.2% when superpixels' borders excluded.
- Note: The ML algorithm replaces all writings, level contours, small roads, by the underlying land use: Not that easy.

Changes in land use 1860-2020

- Roads and railways manually encoded.

  Streets obtained as narrow spaces between built-up using mathematical morphology.

	1860	2020
Built-up	0.57	0.99
Streets	0.29	2.69
Main roads	0.23	0.59
Railways	0.03	0.05
Crops	61.1	42.3
Specialised crops	1.55	2.46
Pastures	8.80	17.8
Meadows	10.4	1.5
Forests	14.9	29.0
Water	2.1	2.6

  - Built-up almost double, roads and streets more than doubled.
  - Crops, the largest land-use, declined by more than a third.
  - Even if specialised crops have increased by more than 60%.
  - Meadows disappeared in favour of pastures, the sum being stable.
  - Forests more than doubled.

Origin of 2020 land-uses (top), Destination of 1860 land-uses (bottom)

	2020	Built-up	Streets	Roads	Railways	Crops	Spe. Crops	Pastures	Meadows	Forests	Water
1860	Built-up	42.33%	0.19%	0.47%	0.25%	0.08%	0.15%	0.28%	0.12%	0.15%	0.15%
	Streets	0.00%	9.74%	0.10%	0.04%	0.02%	0.02%	0.07%	0.01%	0.02%	0.02%
	Roads	0.14%	0.22%	33.33%	0.16%	0.03%	0.04%	0.06%	0.04%	0.07%	0.06%
	Railways	0.02%	0.03%	0.02%	55.51%	0.00%	0.01%	0.01%	0.00%	0.00%	0.01%
	Crops	33.98%	53.09%	30.64%	21.73%	88.10%	57.70%	50.51%	55.90%	34.59%	23.87%
	Spe. Crops	2.79%	4.81%	2.89%	1.68%	0.97%	18.25%	1.04%	0.65%	1.03%	0.84%
	Pastures	4.18%	6.37%	6.18	3.66%%	3.10%	9.18%	19.55%	20.57%	10.50%	7.60%
	Meadows	11.86%	18.81%	19.62%	5.00%	9.61%	23.27%	17.66%	8.70%	14.74%	
	Forests	3.73%	5.32%	5.21%	3.33%	2.33%	3.47%	4.04%	3.76%	43.99%	3.86%
	Water	1.00%	1.45%	1.57%	1.25%	0.38%	1.58%	1.19%	1.30%	0.94%	48.86%

	2020	Built-up	Streets	Roads	Railways	Crops	Spe. Crops	Pastures	Meadows	Forests	Water
1860	Built-up	74.33%	0.88%	0.51%	0.02%	6.03%	0.63%	8.64%	0.30%	7.92%	0.72%
	Streets	0.01%	90.18%	0.21%	0.01%	2.47%	0.17%	4.50%	0.07%	2.18%	0.22%
	Roads	0.50%	2.19%	79.05%	0.03%	5.08%	0.41%	4.23%	0.19%	7.79%	0.55%
	Railways	0.70%	2.49%	0.37%	83.74%	2.85%	0.87%	4.75%	0.07%	3.55%	0.61%
	Crops	0.55%	2.33%	0.32%	0.02%	61.02%	2.32%	14.67%	1.32%	16.43%	1.03%
	Spe. crops	1.79%	8.33%	1.19%	0.06%	26.52%	28.96%	11.91%	0.60%	19.27%	1.42%
	Pastures	0.47%	1.95%	0.45%	0.02%	14.88%	2.57%	39.41%	3.38%	34.62%	2.28%
	Meadows	1.13%	4.86%	1.20%	0.06%	20.34%	2.27%	39.71%	2.45%	24.29%	3.74%
	Forests	0.25%	0.96%	0.22%	0.01%	6.60%	0.57%	4.80%	0.36%	85.55%	0.68%
	Water	0.48%	1.90%	0.49%	0.03%	7.78%	1.90%	10.27%	0.92%	13.33%	62.93%

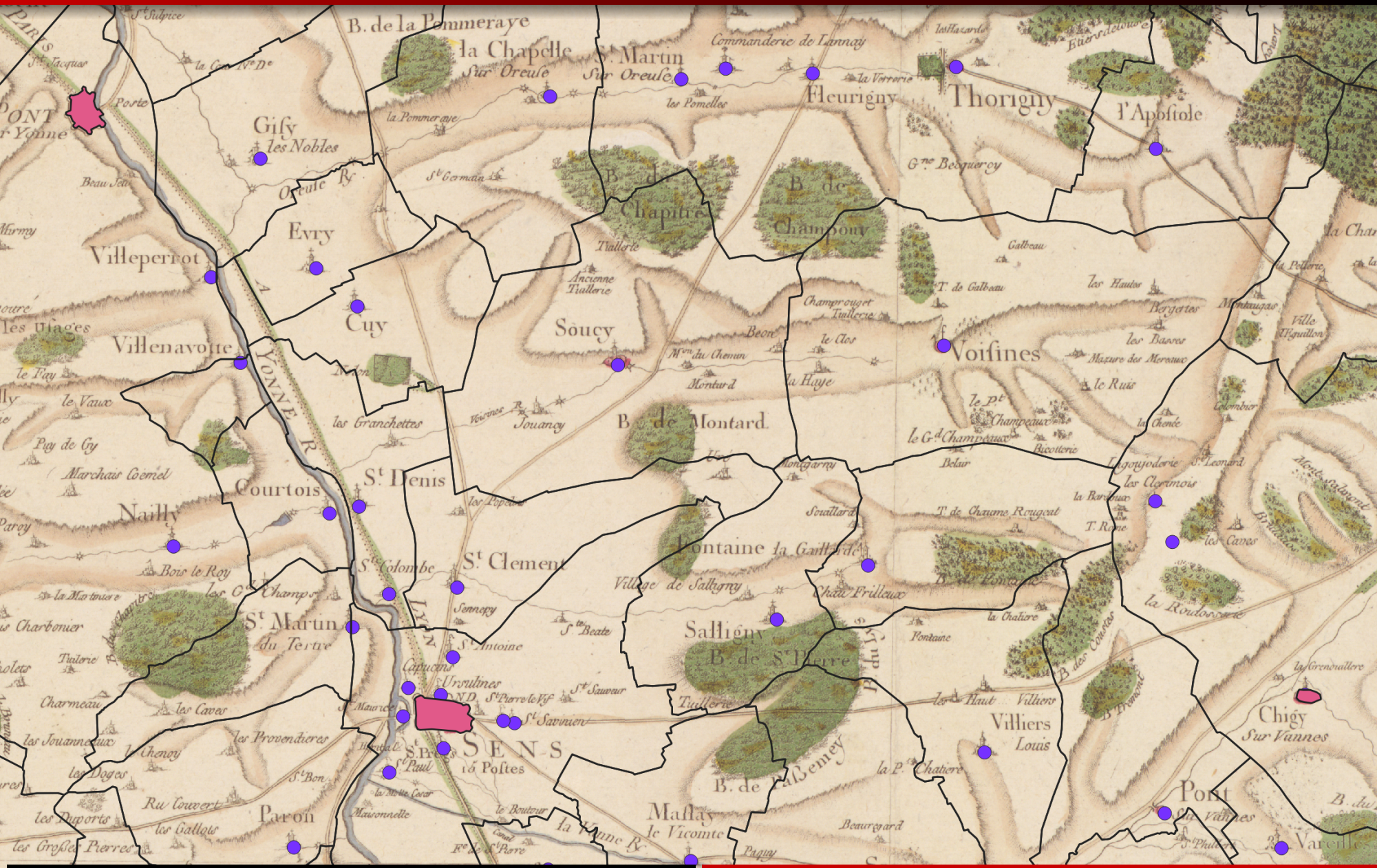
- 2020 Built-up/roads/streets grew from crops/pastures, <4% 1860 crops became built-up/road/streets.
- 31% crops→pastures/forests, 34.6% pastures→forests, 65% meadows→pastures/forests.

Data for other points in time

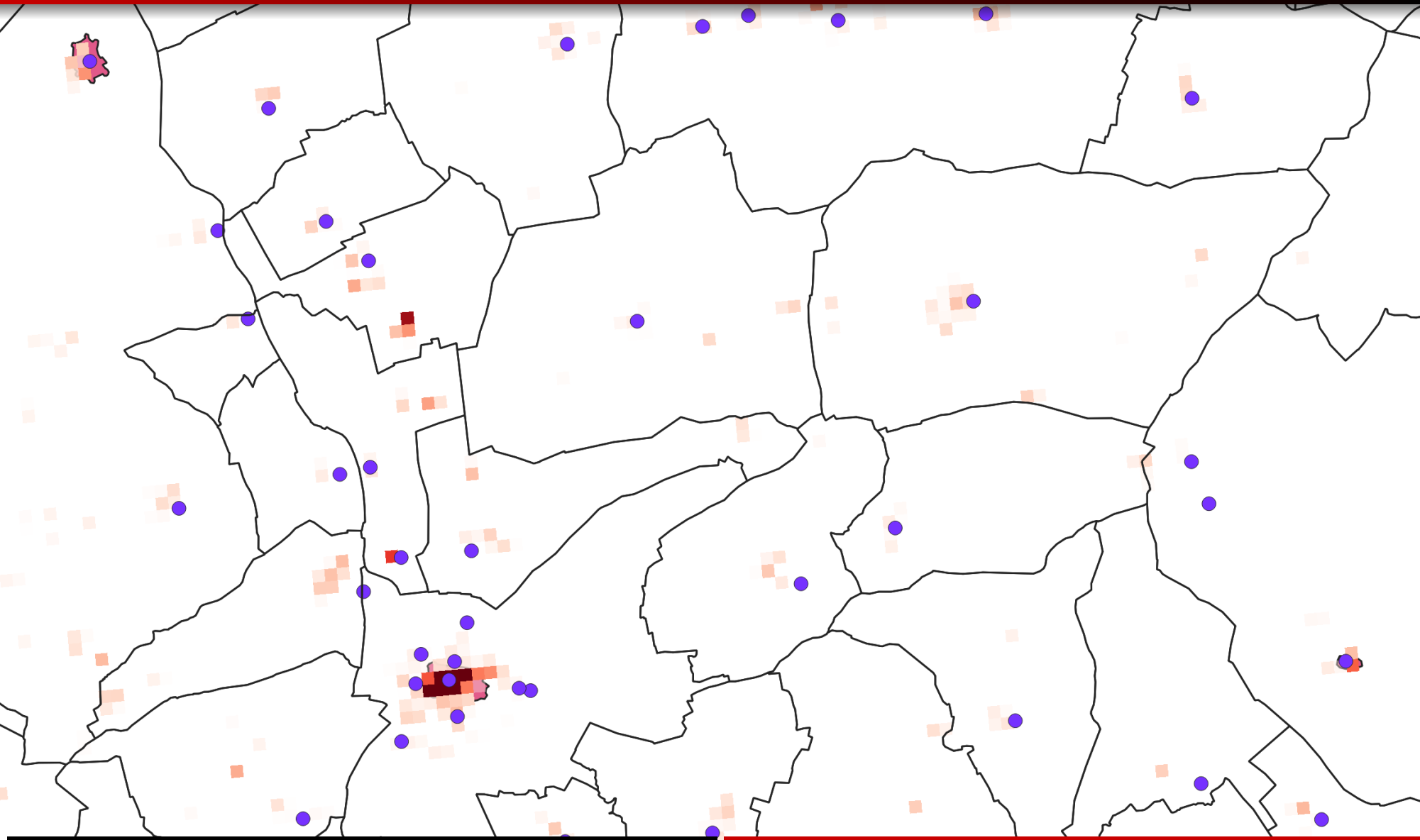
- Cassini maps (1760):
  - Much simpler, one could apply the same kind of strategy.
  - But details are given only for areas with a large population, symbols otherwise, and the purpose is to have built-up information similar to other maps.
  - Manually encoded as the (point) symbols (churches, castles,...).
  - We also have information about buildings built before 18^{th} c. that still exist.
  - We use all of that within another random forest to predict 1760 built-up even outside urban areas represented on the maps.

$\Rightarrow$ More similar in nature to 1860 and 2020 information but very small villages/isolated buildings are still missing.

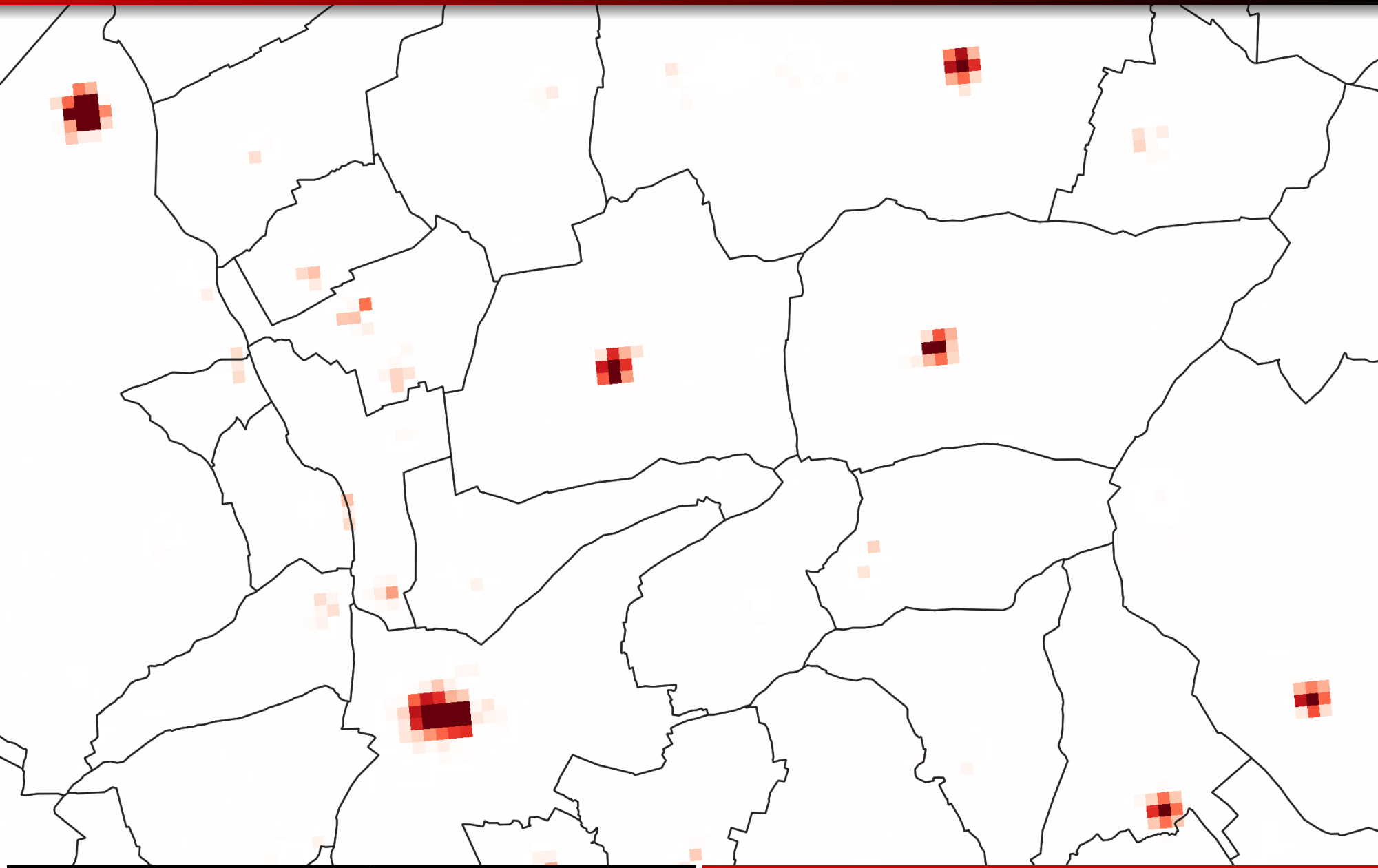
Raw Cassini maps, Sens (Burgundy)



Cassini built-up and symbols and 18^{th} c. remaining built-up, Sens (Burgundy)



1760 built-up prediction by random forest, Sens (Burgundy)

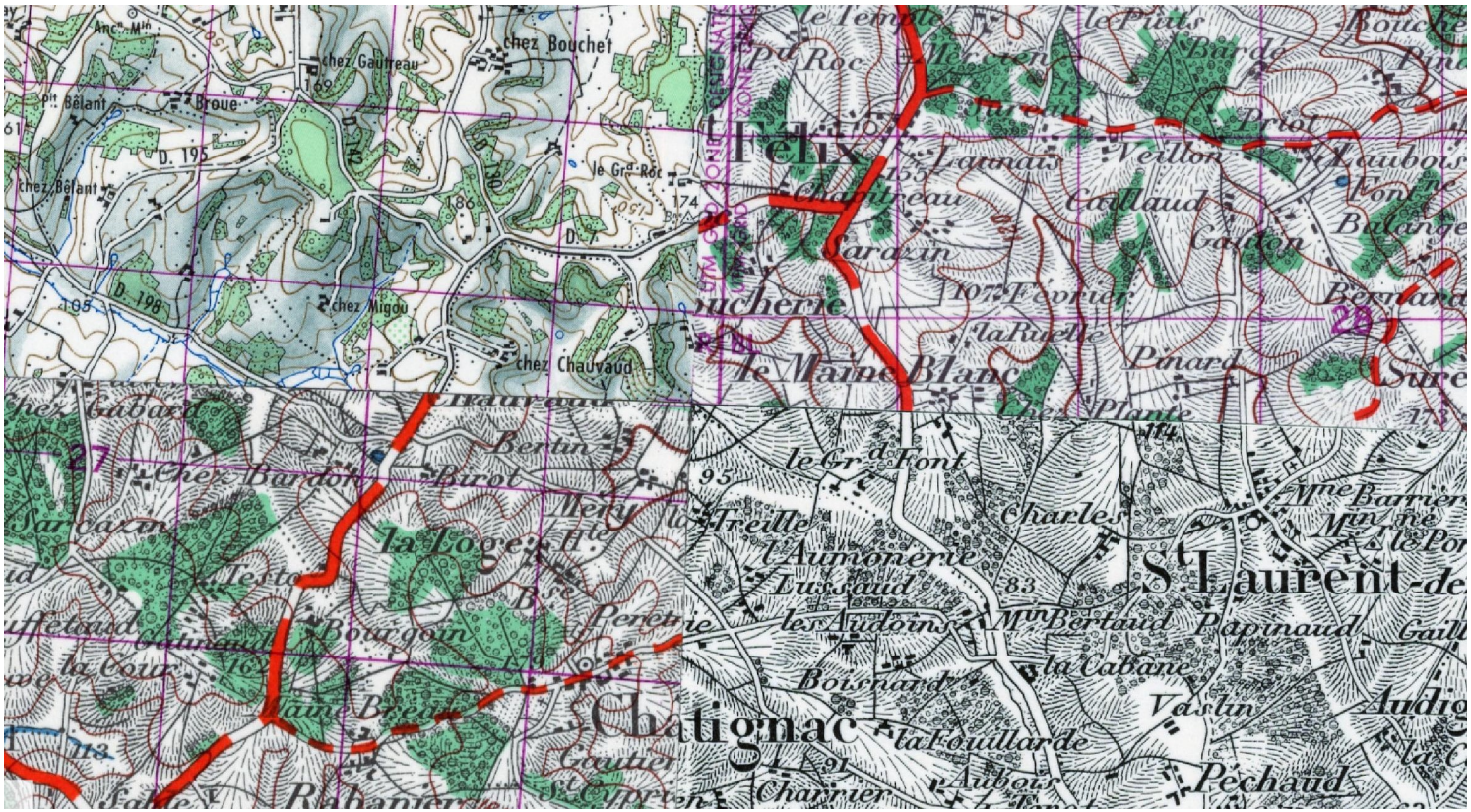


1960 Built-up, U-Net Convolutional Neural Network

- 1960 maps: Little use of colours, typically buildings in black as all writings.

⇒ Requires more powerful ML methods: Convolutional Neural Networks.

- Also solves for another issue: 5 different types of legends (way of representing buildings, forests, roads...) over France, which even changes within tiles.



1960 map, Marseille

1960 built-up, Marseille

Population data

- Historical population censuses:
  - Since 1793, every 5/10 years, at the municipality level (c. 36,000 units).
  - Cassini project's crosswalk: Lists all mergers and splits of municipalities since 1793.
  - We developed an algorithm to attribute municipality boundaries at each census date consistent with the 2020 municipality boundaries.
- We create a 200m×200m population gridded data set for 1760, 1860, 1960 and 2020 that:
  - Aggregates 4m×4m pixel information for land use,
  - Allocates municipal population to pixels proportionally to total built-up, as done in various current projects at the word level (GHS Pop, World Pop).
  - Variant that uses a predicted height of built-up based on buildings from that time that still exist.

Delineation of cities throughout history

- Purpose: Use of a common (statistical) methodology to delineate cities (metropolitan areas) at each date to compare their spatial extent and expansion over time in a meaningful way.
- Standard approach: Absolute thresholds constant over time.
  Eg Bairoch's cities: Aggregates of urban municipalities (larger than 2,000 inhabitants) with, overall, more than 5,000 inhabitants at any date.
- Feature: Almost by construction, increasing number of cities over time.
  191 Bairoch's cities in 1760, 339 in 1860, and $\approx$500 nowadays.
- We use a strategy with relative thresholds, which are local and year-specific.

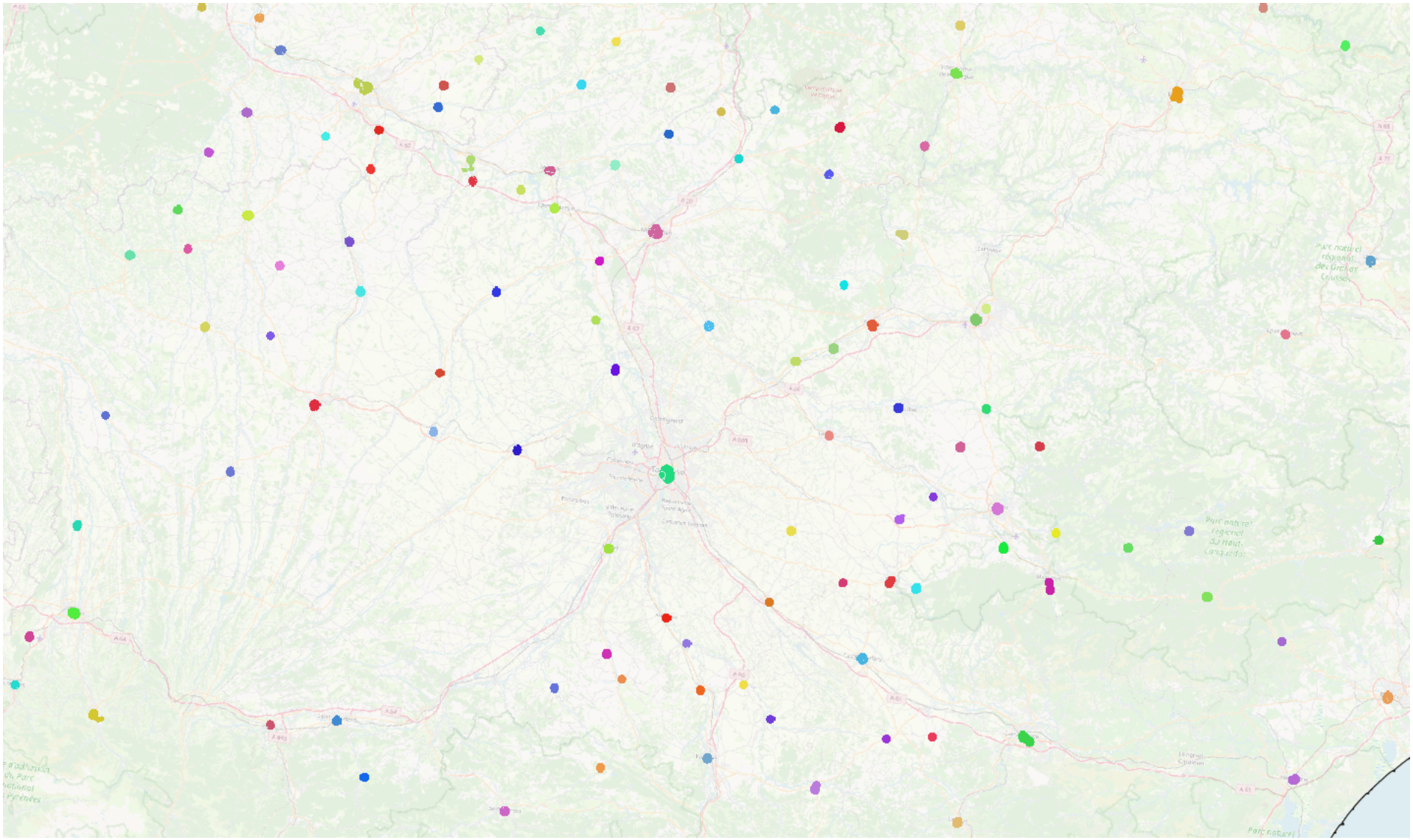
$\Rightarrow$ Urbanisation can be studied in a consistent way over the long run allowing for the emergence and disappearance of cities.

Delineation of cities throughout history

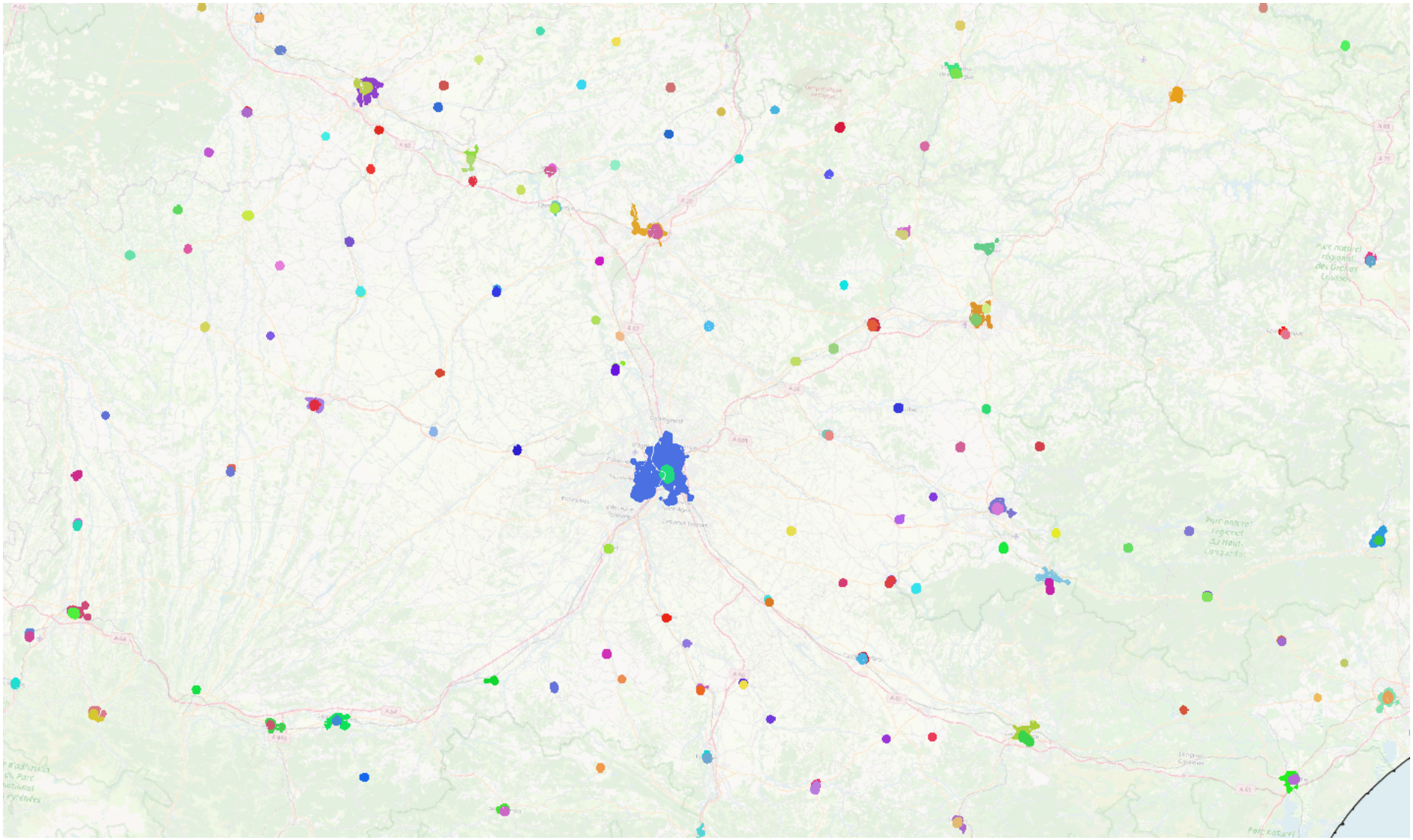
- Methodology: De Bellefon, Combes, Duranton, Gobillon, and Gorin (2021):
  - Computes the gridded (smoothed) population density for the whole of France.
  - 5,000 random reshuffles of all populated pixels over all (livable) pixels:

  $\Rightarrow$ Counterfactual building density (smoothed) distribution for each pixel under randomness.
  - Livable pixels: Below 99^{th} percentile of built pixels for elevation, slope, water.
  - Urban pixels: Those where (smoothed) density is above the 95^{th} percentile of the (smoothed) counterfactual density distribution.
  - Urban areas: Sets of contiguous urban pixels.
  - Re-shuffling repeated within urban areas only: Urban pixels at second order, contiguous ones correspond to 'urban cores'.
  - Cities: Urban areas with at least one urban core.
  - Note: This is a relative definition of cities ('significant peaks of population density')

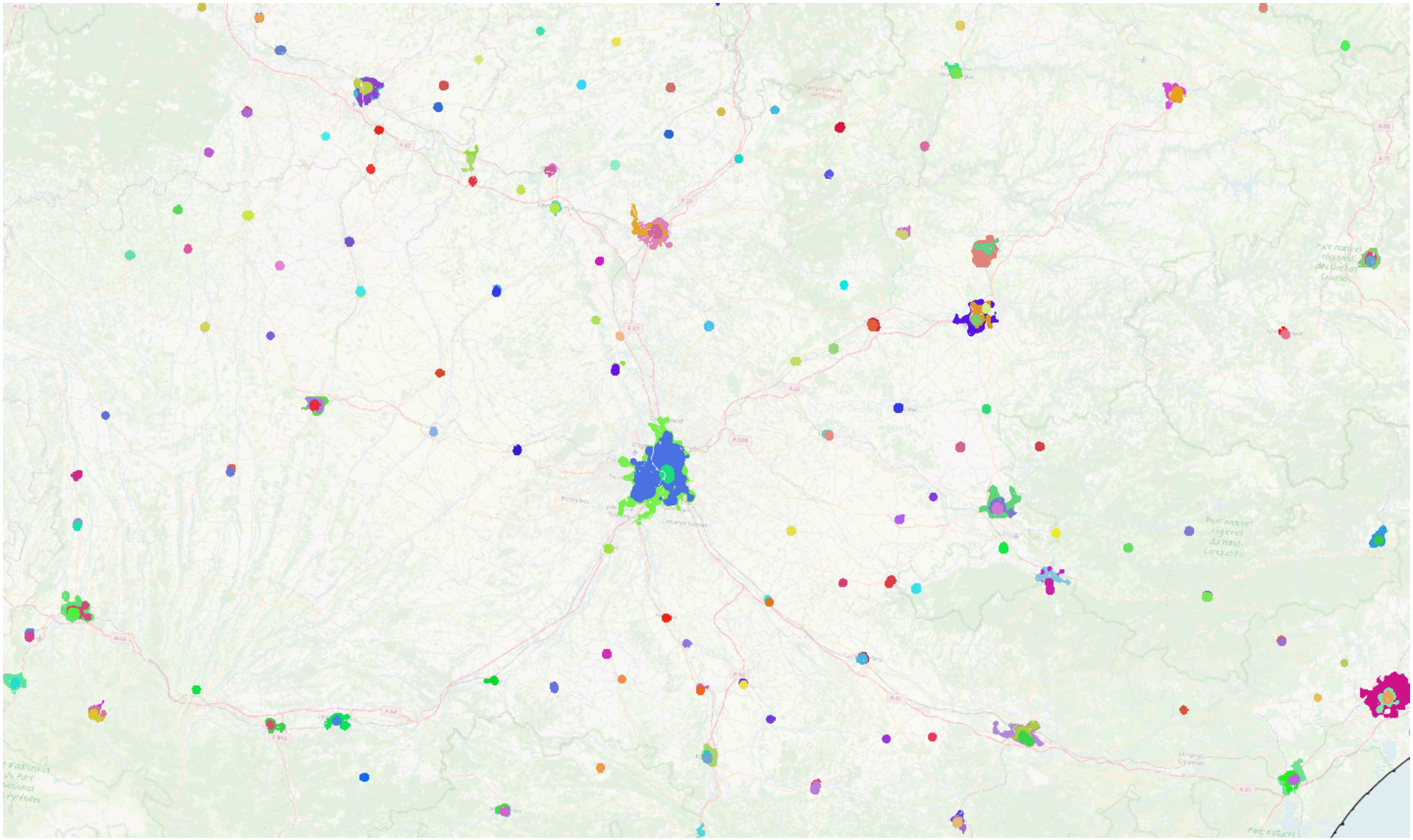
1760 Cities around Toulouse



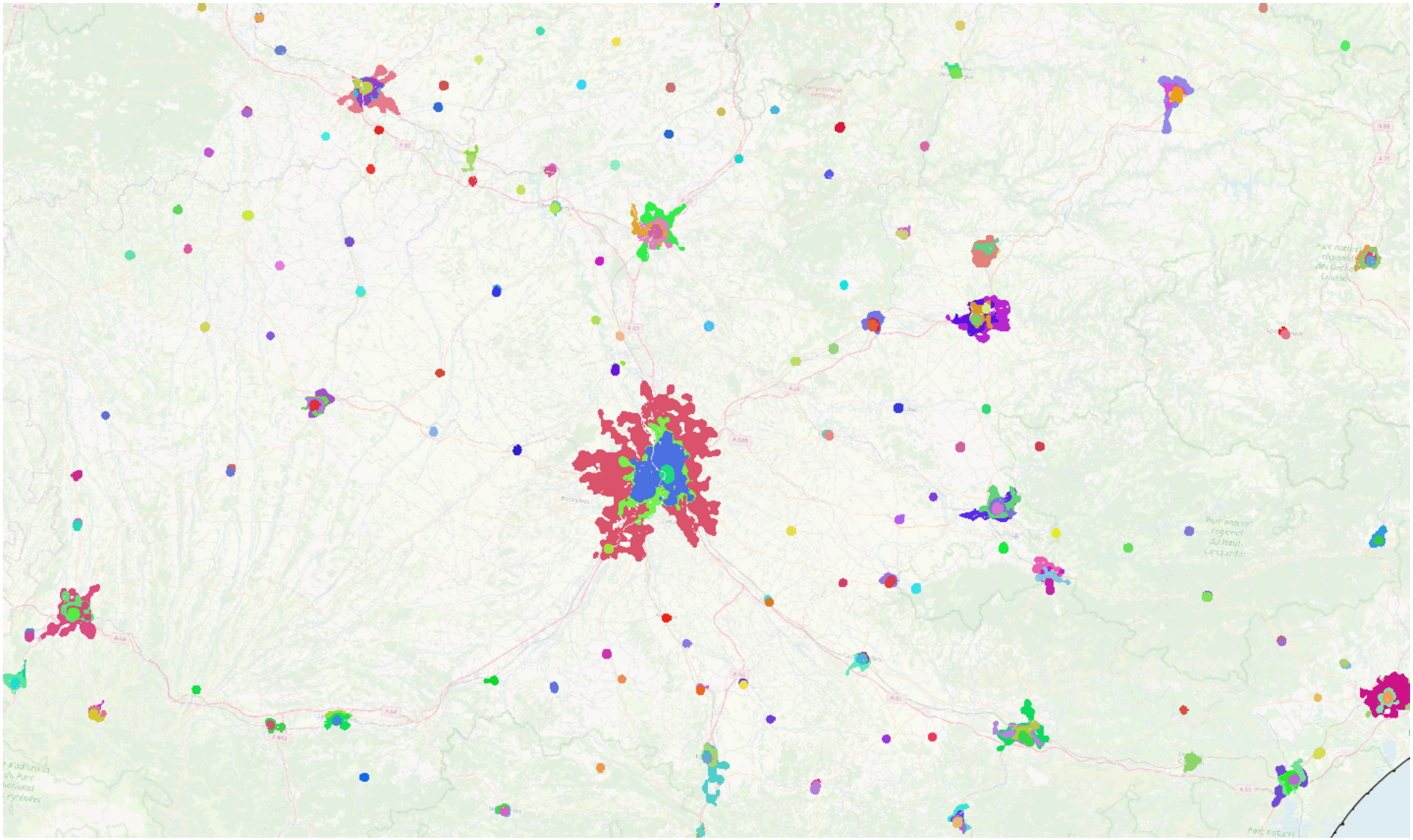
1760 and 1860 Cities around Toulouse



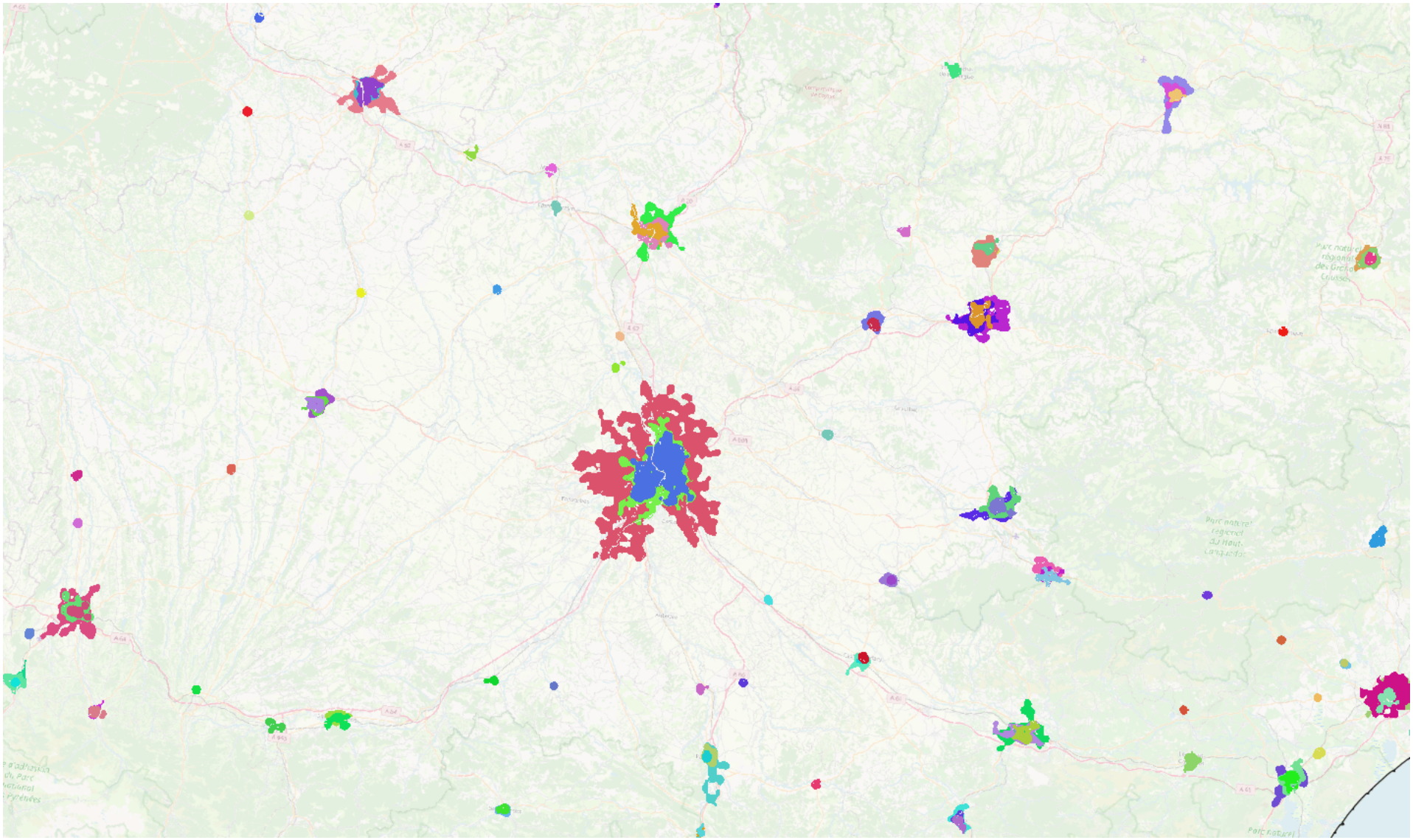
1760, 1860 and 1960 Cities around Toulouse



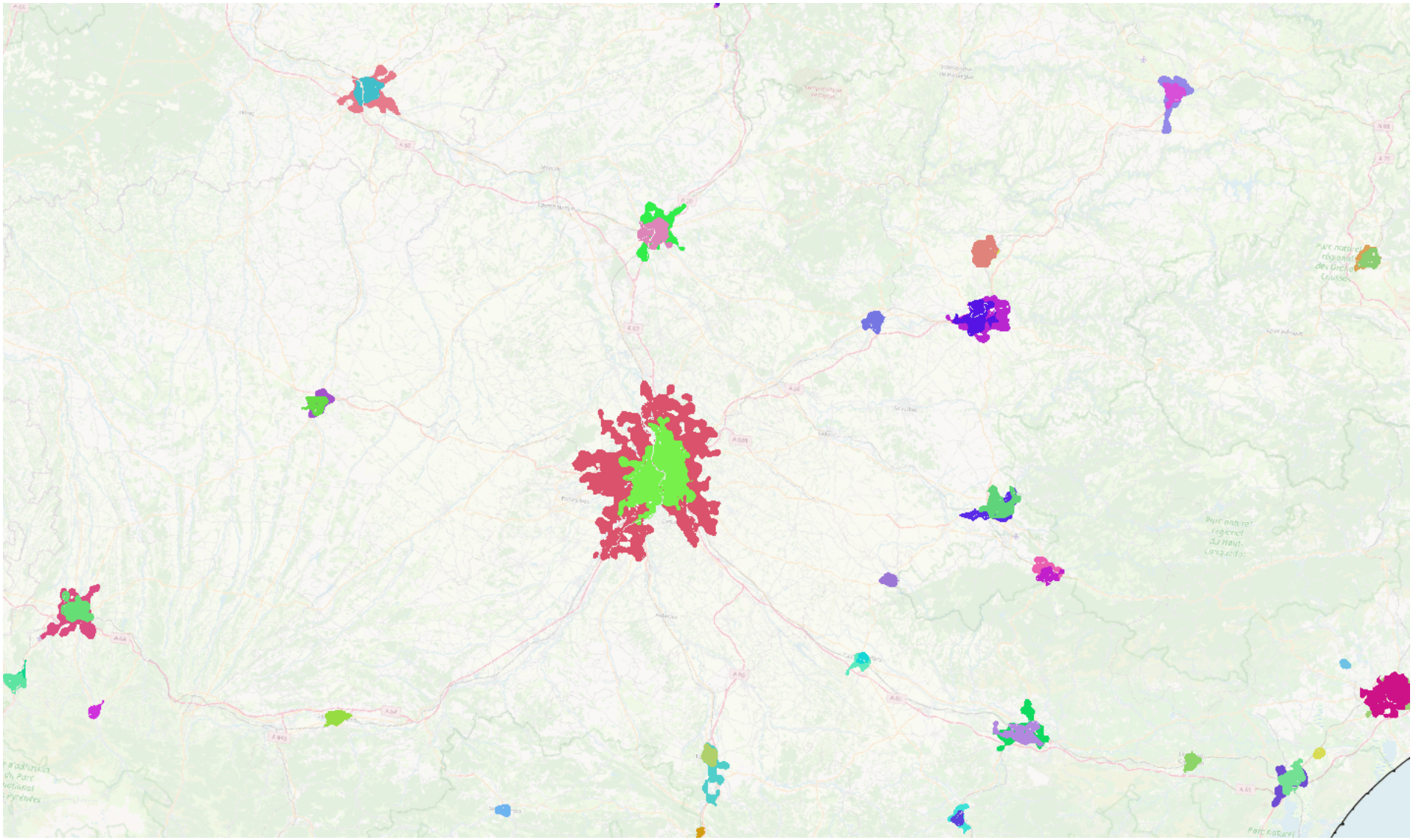
1760, 1860, 1960 and 2020 Cities around Toulouse



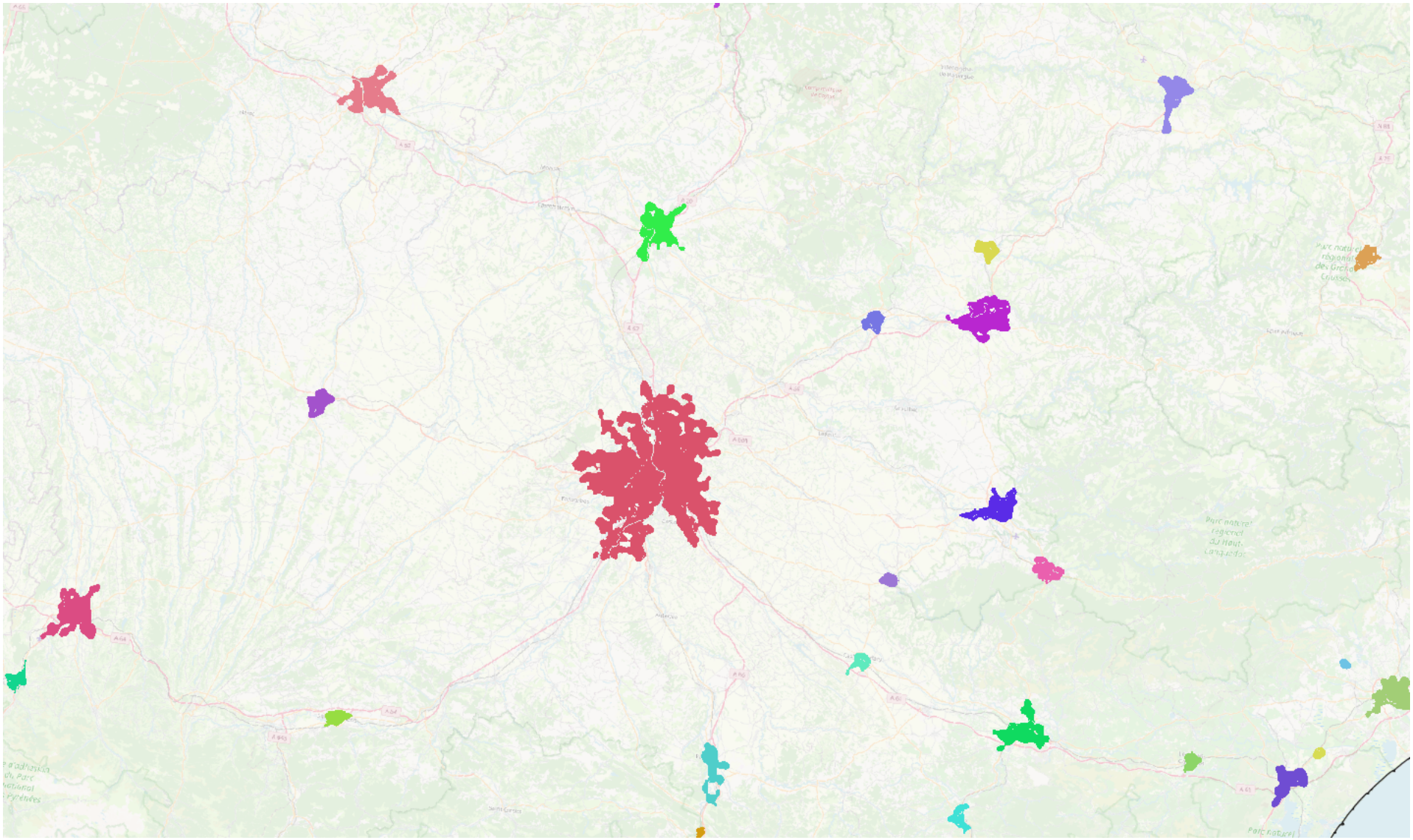
1860, 1960 and 2020 Cities around Toulouse



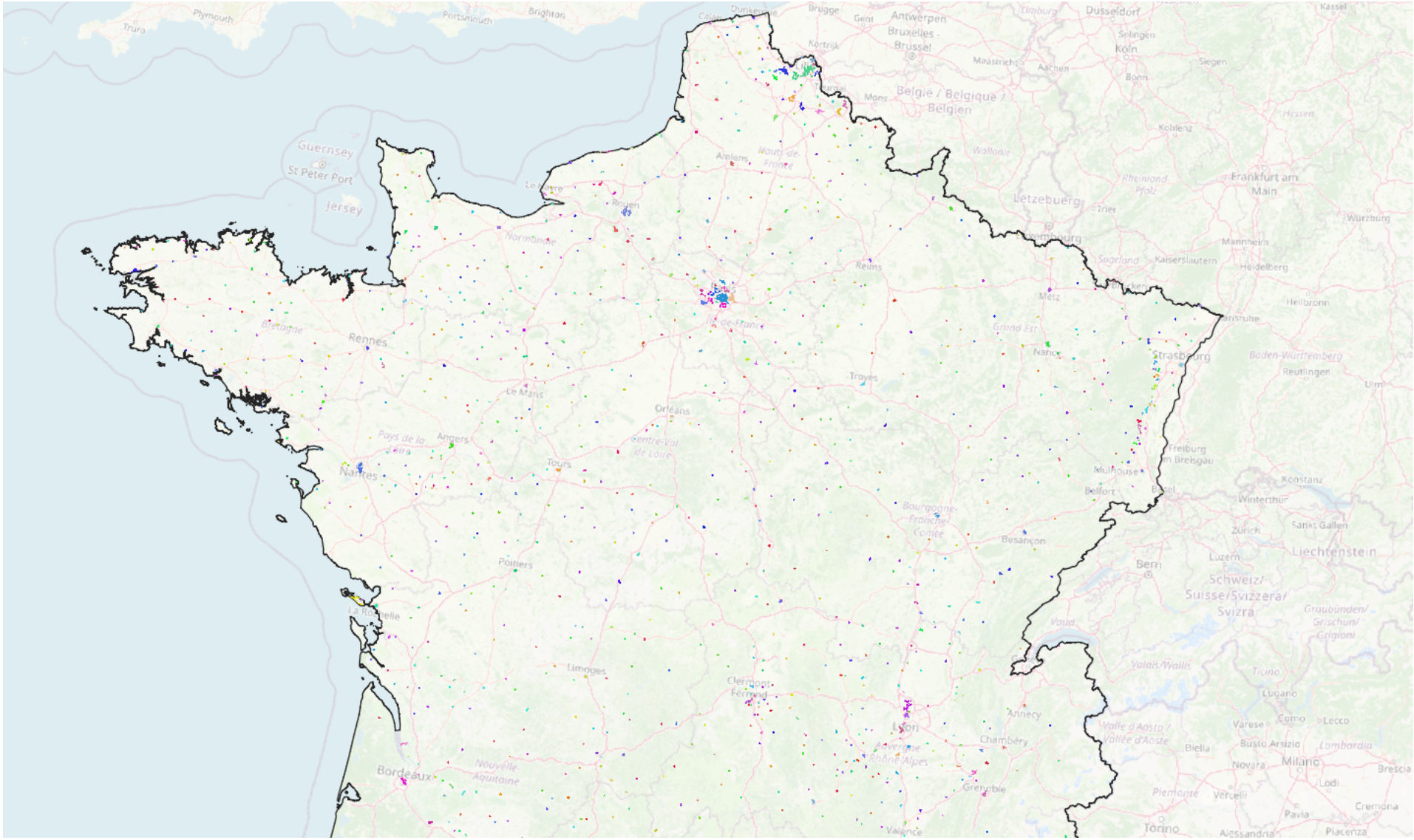
1960 and 2020 Cities around Toulouse



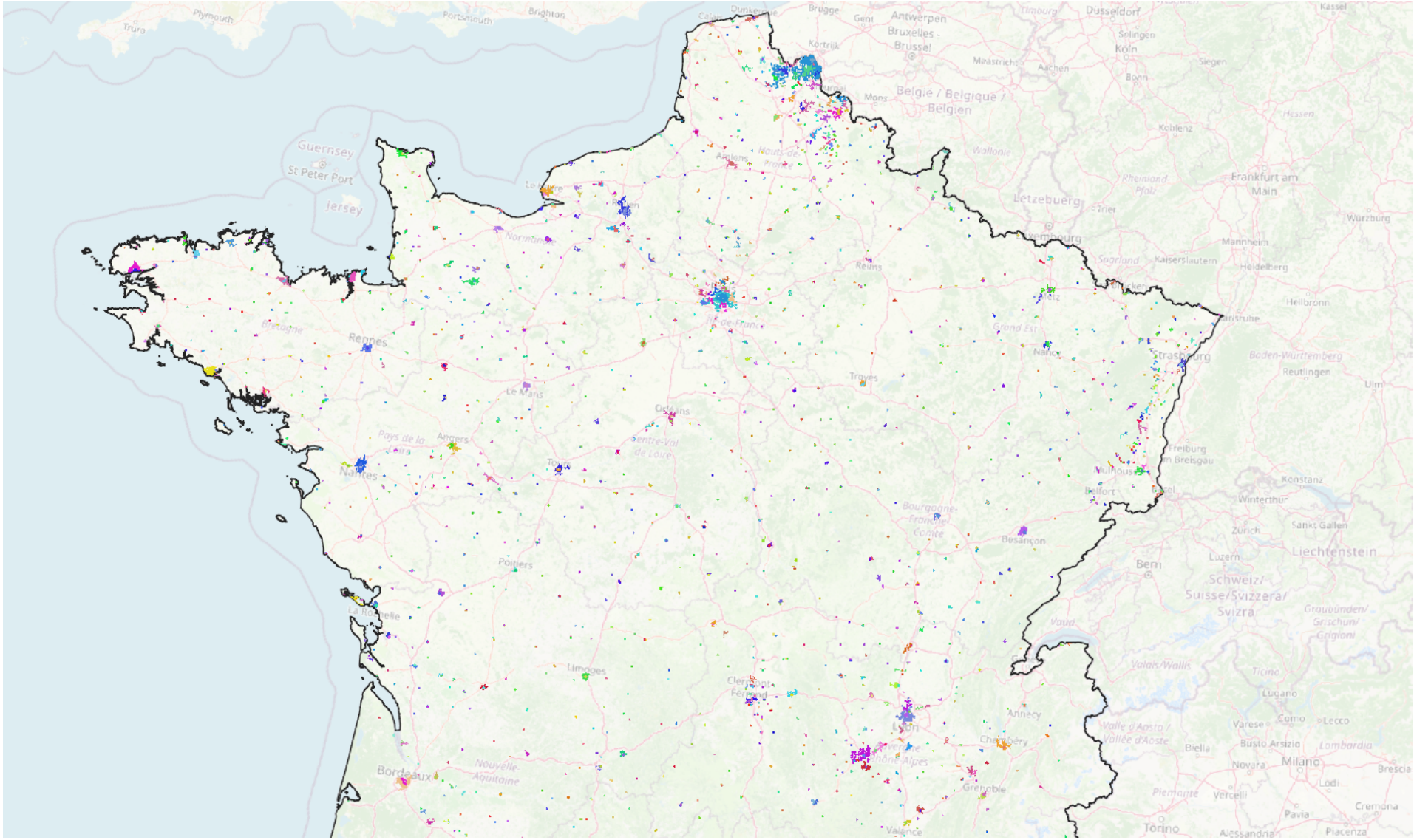
2020 Cities around Toulouse



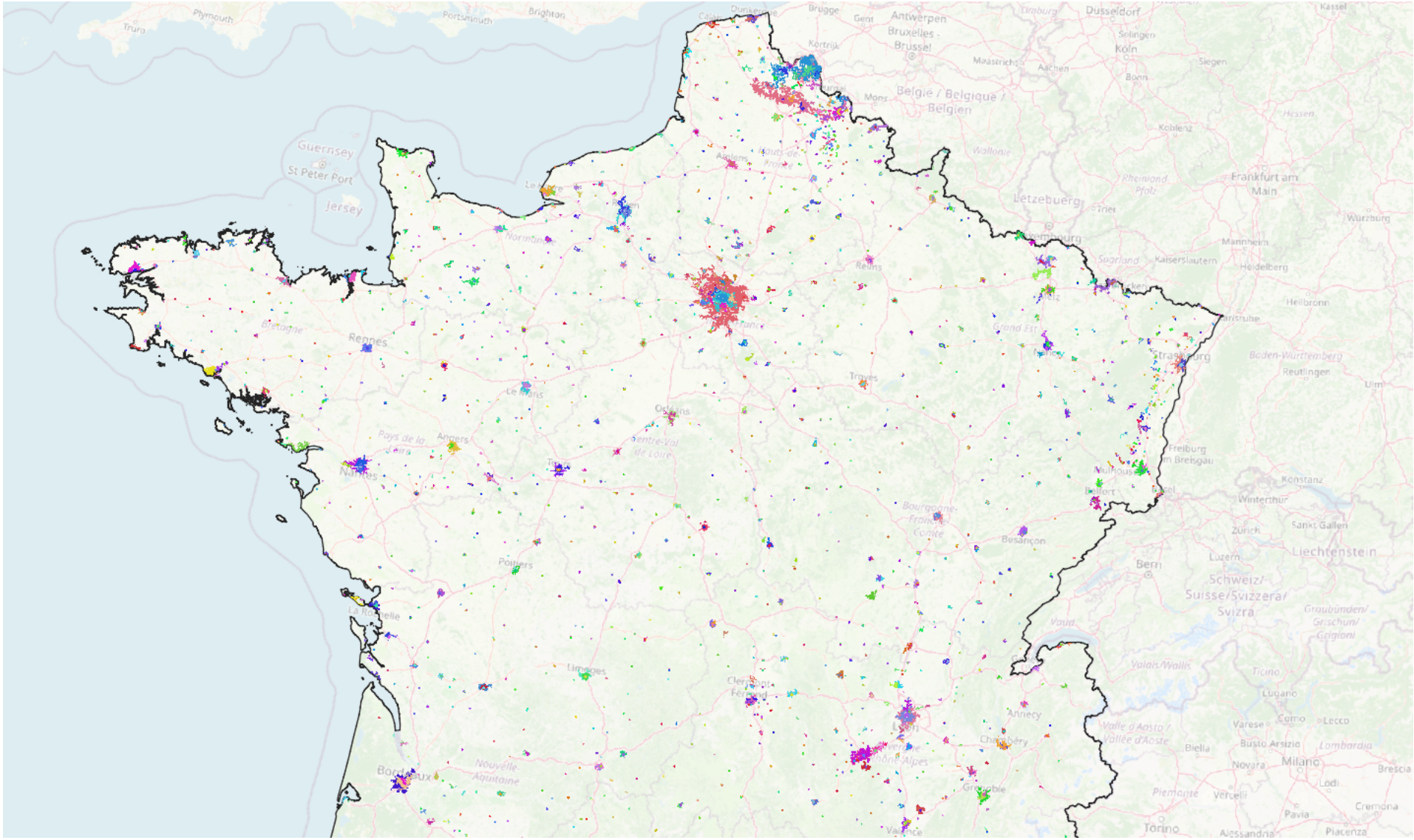
1760 Cities


Plymouth
Truro
Portsmouth
Brighton
Guernsey
St Peter Port
Jersey
Brugge
Gent
Antwerpen
Bruxelles - Brussel
België / Belgique / Belgien
Maastricht
Aachen
Köln
Bonn
Siegen
Hessen
Koblenz
Wallonie
Amiens
Hauts-de-France
Rouen
Normandie
Reims
Luxembourg
Trier
Rheinland-Pfalz
Saarland
Kaiserslautern
Mannheim
Heidelberg
Frankfurt am Main
Würzburg
Heilbronn
Karlsruhe
Baden-Württemberg
Reutlingen
Ulm
Bretagne
Rennes
Le Mans
Île-de-France
Paris
Troyes
Metz
Nancy
Grand Est
Strasbourg
Freiburg im Breisgau
Mulhouse
Belfort
Basel
Konstanz
Winterthur
Zürich
Sankt Gallen
Liechtenstein
Pays de la Loire
Angers
Nantes
Orléans
Tours
Centre-Val de Loire
Bourgogne-Franche-Comté
Besançon
Poitiers
La Rochelle
Limoges
Clermont-Ferrand
Bern
Luzern
Schweiz/Suisse/Svizzera/Svizra
Graubünden/Grischun/Grigioni
Vaud
Valais/Wallis
Ticino
Lugano
Como
Lecco
Varese
Busto Arsizio
Lombardia
Milano
Brescia
Lodi
Cremona
Pavia
Piacenza
Torino
Alessandria
Vercelli
Piemonte
Valle d'Aosta / Vallée d'Aoste
Biella
Novara
Lyon
Annecy
Chambéry
Auvergne-Rhône-Alpes
Grenoble
Valence
Nouvelle-Aquitaine
Bordeaux

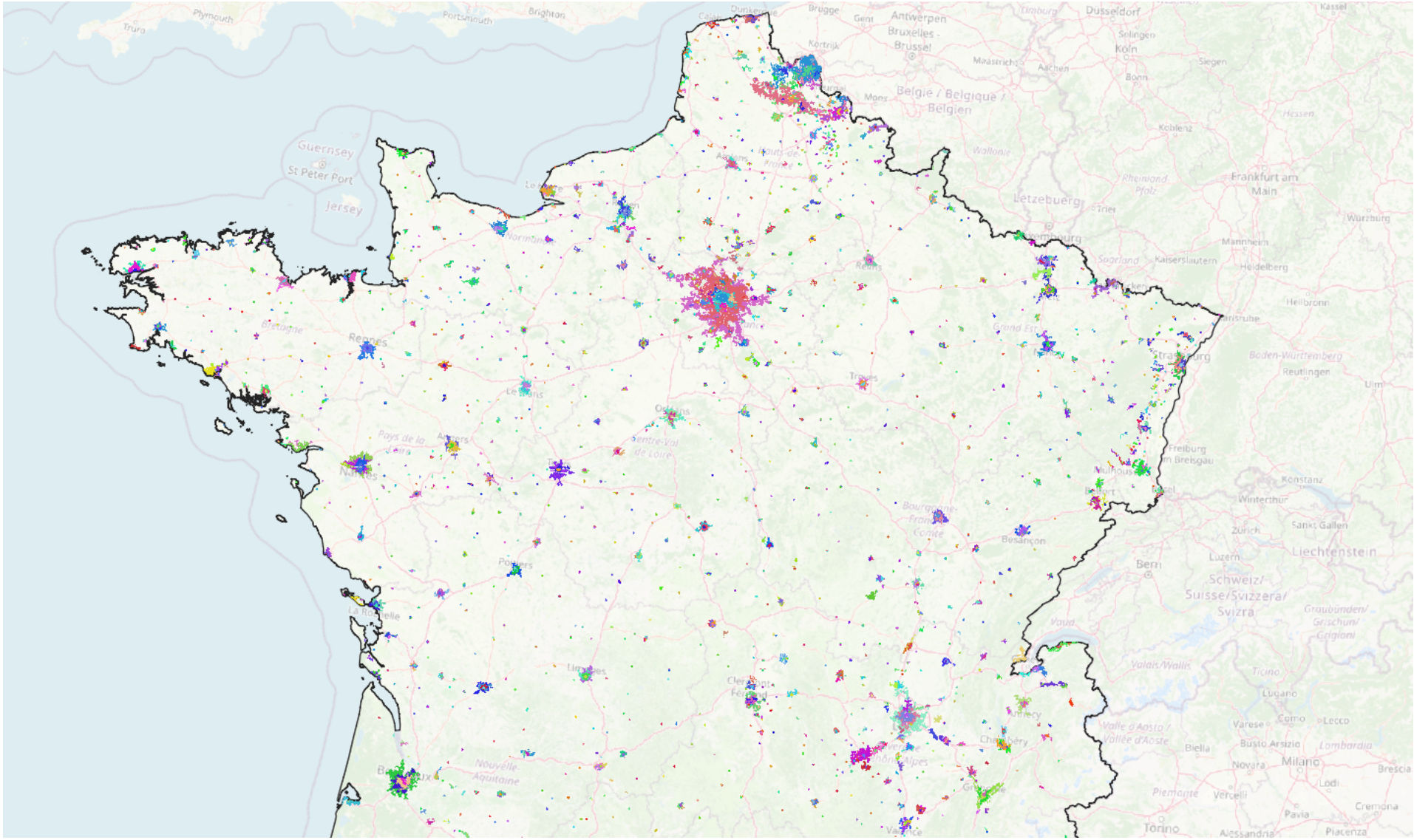
1760 and 1860 Cities



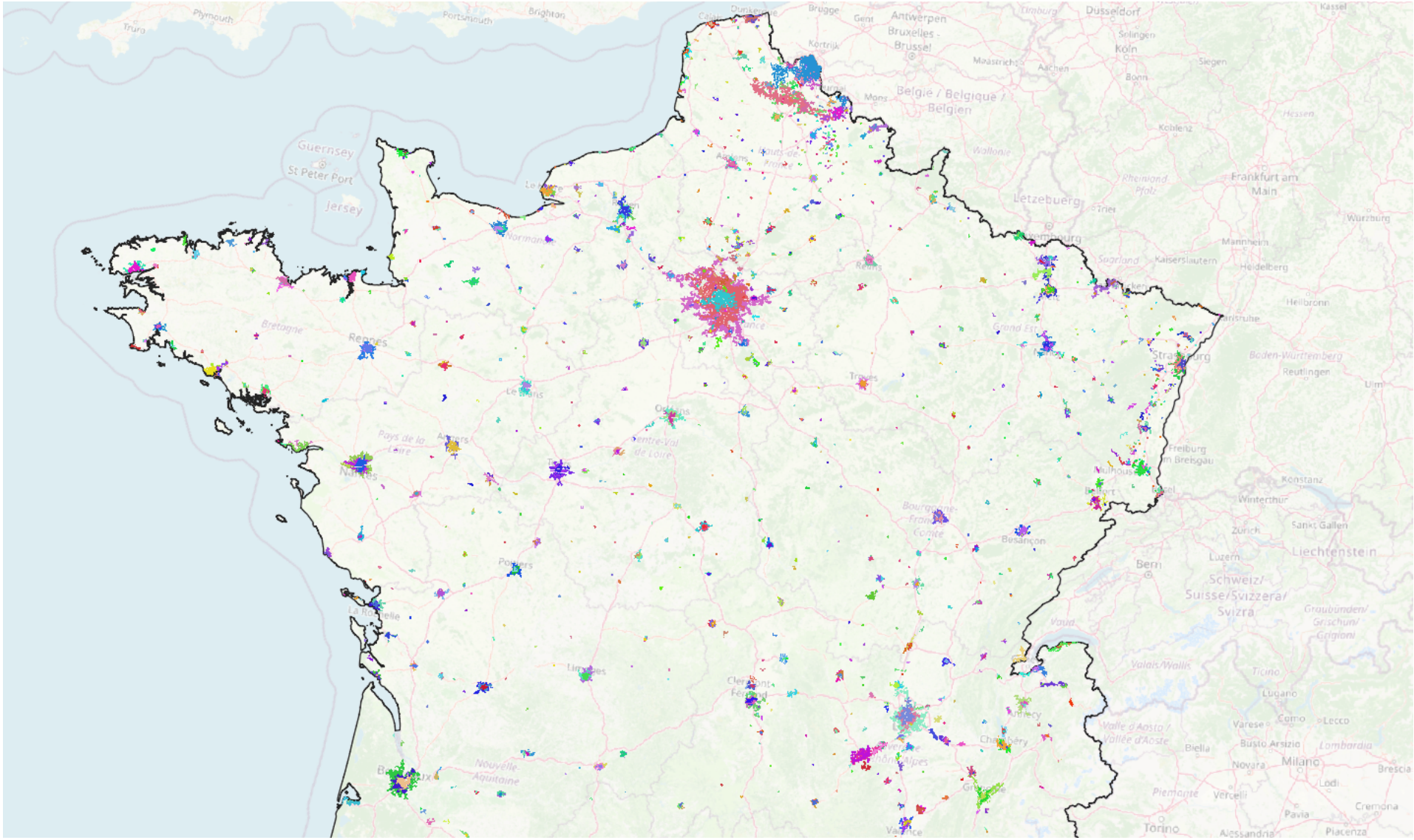
1760, 1860, 1960 and 2020 Cities



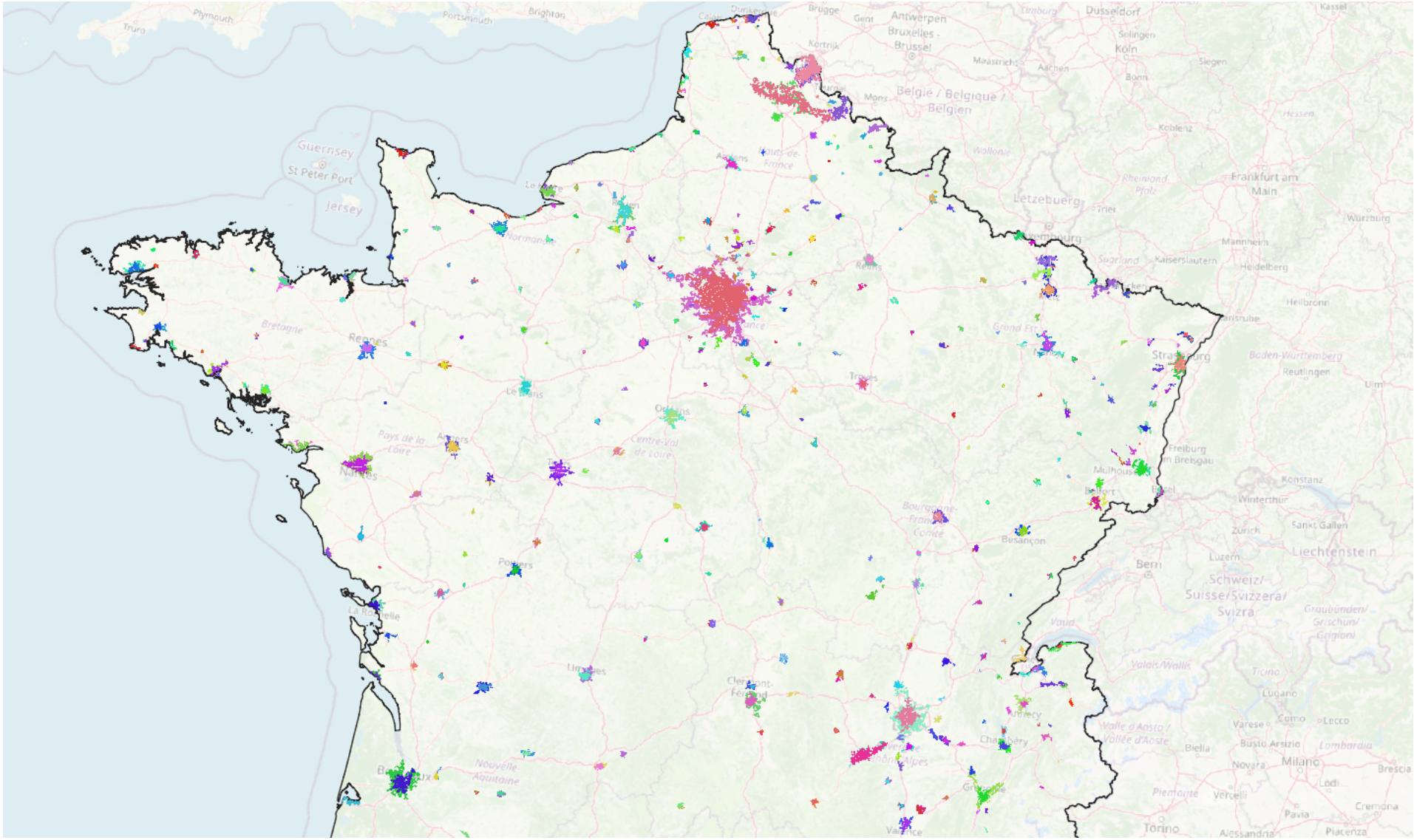
1760, 1860, 1960 and 2020 Cities



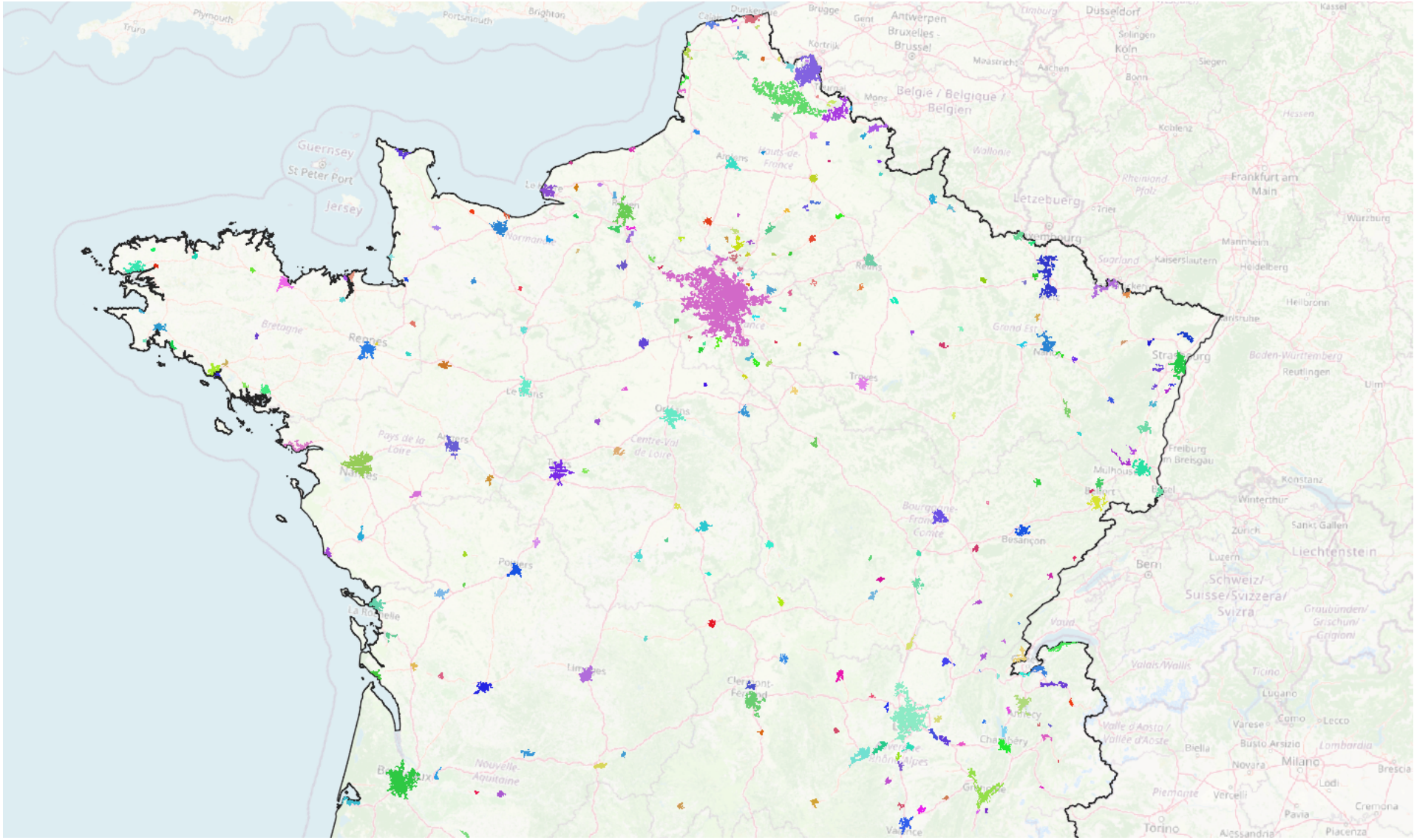
1860, 1960 and 2020 Cities



1960 and 2020 Cities



2020 Cities


Plymouth
Truro
Portsmouth
Brighton
Guernsey
St Peter Port
Jersey
Brugge
Gent
Antwerpen
Bruxelles - Brussel
Kortrijk
Dunkerque
Calais
Maastricht
Aachen
Düsseldorf
Solingen
Köln
Bonn
Siegen
Mons
België / Belgique / Belgien
Koblenz
Hessen
Wallonie
Hauts-de-France
Rheinland-Pfalz
Frankfurt am Main
Letzebuerg
Trier
Würzburg
Mannheim
Saarland
Kaiserslautern
Heidelberg
Heilbronn
Karlsruhe
Grand Est
Baden-Württemberg
Reutlingen
Ulm
Bretagne
Rennes
Le Mans
Troyes
Orléans
Pays de la Loire
Nantes
Centre-Val de Loire
Freiburg im Breisgau
Mulhouse
Konstanz
Winterthur
Zürich
Sankt Gallen
Liechtenstein
Bourgogne-Franche-Comté
Besançon
Bern
Luzern
Schweiz/Suisse/Svizzera/Svizra
Graubünden/Grischun/Grigioni
Poitiers
La Rochelle
Limoges
Vaud
Valais/Wallis
Ticino
Lugano
Varese
Como
Lecco
Busto Arsizio
Lombardia
Milano
Novara
Brescia
Lodi
Piemonte
Vercelli
Cremona
Pavia
Torino
Alessandria
Piacenza
Nouvelle-Aquitaine
Bordeaux
Clermont-Ferrand
Auvergne-Rhône-Alpes
Annecy
Chambéry
Valle d'Aosta / Vallée d'Aoste
Grenoble
Valence

Urbanisation and cities over time

Year	French Population	Number	Urban Population	Urban Pop. Share	Urban Area Share	Urban pop. Density
1760	28.13	1622	6.78	24.1	1.3	917.8
1860	36.75	979	9.90	26.9	1.6	1100.8
1960	44.38	337	22.57	50.8	2.0	2068.9
2020	65.71	382	37.94	57.7	3.4	2011.9

Shares and growth rates in %.

- Large increase in urbanisation (population and area)
- Large decrease in the number of cities,
- Density doubled, mostly over 1860-1960.

Decomposition of urbanisation changes at the country level

- Variations over time of urbanisation can be due to:
  - Changes in thresholds (counterfactual density under randomness),
  - Changes in density at given thresholds.
- Possible to compare density in a given year to thresholds of another year.
- For instance for the 1860/2020 comparison, four types of urban pixels:
  - Urban pixels in 1860 according to 1860 thresholds,
  - Urban pixels in 2020 according to 2020 thresholds,

But also

  - Urban pixels in 1860 according to 2020 thresholds,
  - Urban pixels in 2020 according to 1860 thresholds.

Conditional urbanisation rates

	Thresholds											
	1760			1860			1960			2020		
Year	#	% Pop.	% Area	#	% Pop.	% Area	#	% Pop.	% Area	#	% Pop.	% Area
1760	1622	24.1	1.3	952	19.7	0.7	462	15.1	0.3	356	13.8	0.2
1860	2414	34.9	3.2	979	26.9	1.6	352	20.6	0.7	240	18.5	0.5
1960	1379	61.5	4.4	796	57.5	3.2	337	50.8	2.0	244	47.7	1.5
2020	3617	78.5	10.1	2182	73.9	7.8	783	64.1	4.8	382	57.7	3.4

- Using next period's thresholds much reduces the number of previous period's cities.

  For instance, relatively constant number of cities using 2020 thresholds; still urban population share multiplied by 4 and area by 15.

- Reversely, using previous period's thresholds makes the number of cities increase, and much increases urbanisation rates.

$\Rightarrow$ Absolute contemporaneous thresholds hide a large chunk of urbanisation.

2/3 of cities are missing, urbanisation rates underestimated by more than a third.

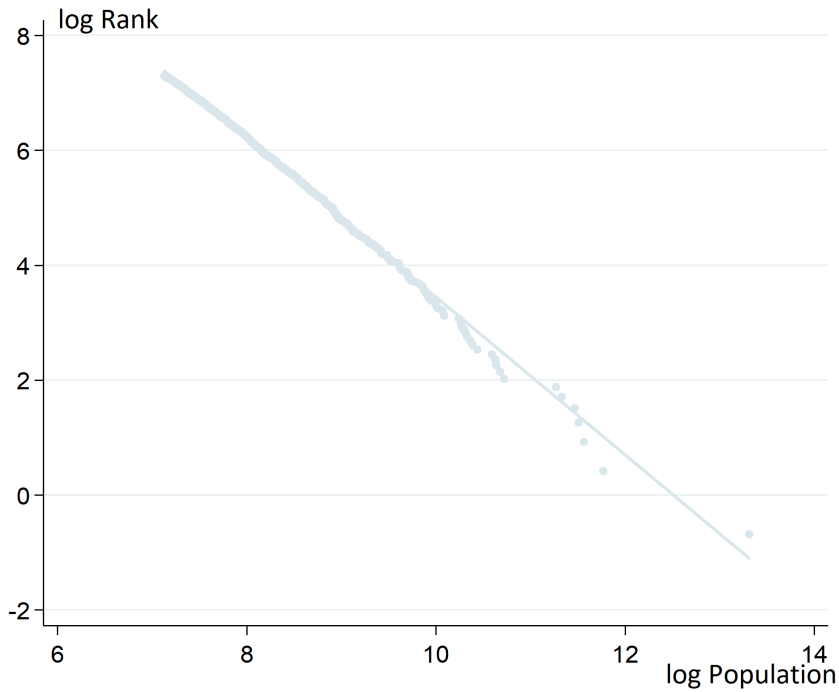
City size distribution: Zipf's Law over history

- At any date, cities of very different sizes co-exist.
- Zipf's law:
  - Let Pop_c be the population of city c and R_c its rank.
  - Estimate

$$\log\left(R_c - 1/2\right) = \alpha - \beta \log Pop_c + \varepsilon_c.$$

$\Rightarrow$ β expected to be close to 1.

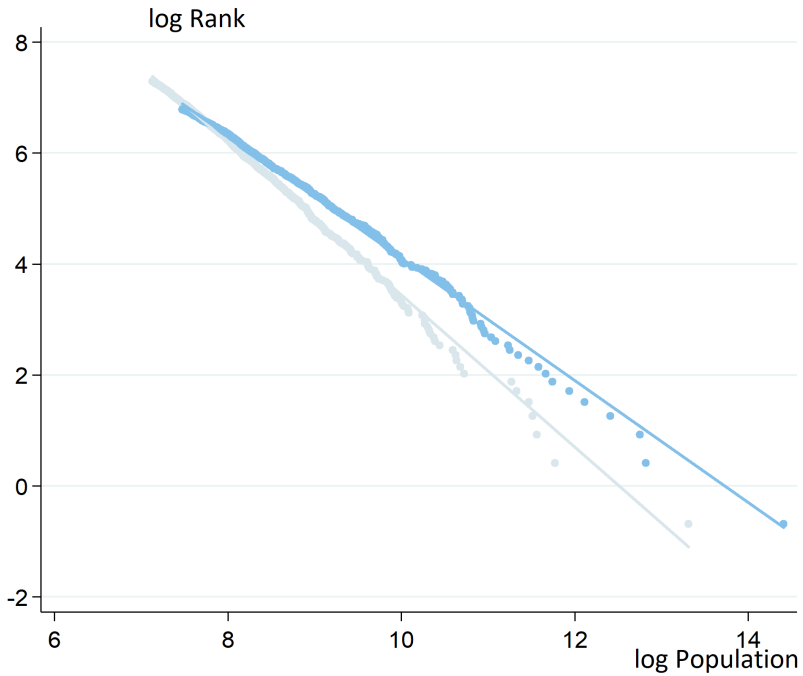
City size distribution: 1760



	All cities			10% smallest excluded			100 Largest		
Year	β	R^2	N	β	R^2	N	β	R^2	N
1760	-1.32^{***}	0.98	1,622	-1.37^{***}	1.00	1,459	-1.42^{***}	0.99	100

- Very good fit, large slope (cities are less unevenly distributed than predicted by Zipf's law).
- Paris larger than its predicted value.

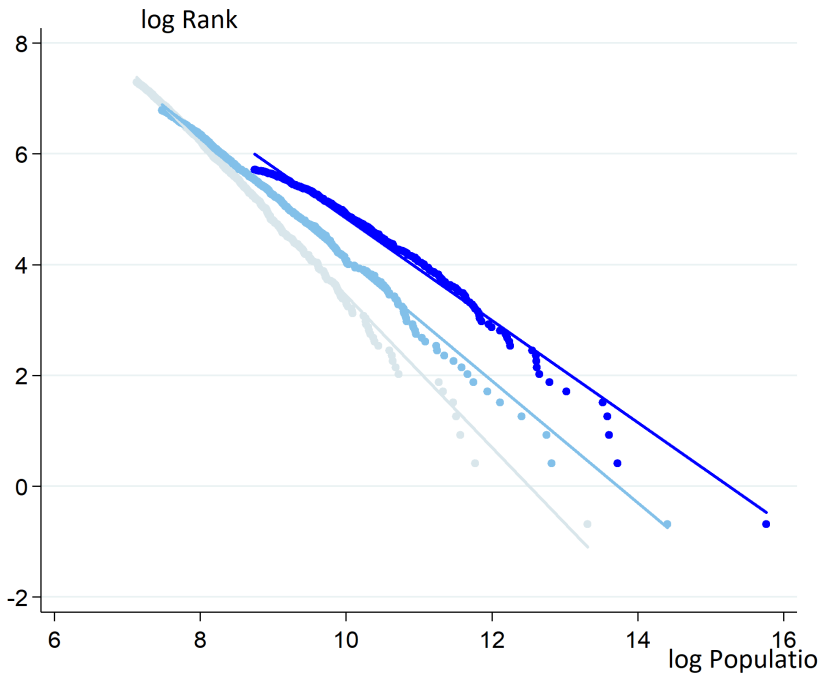
City size distribution: 1760 and 1860



	All cities			10% smallest excluded			100 Largest		
Year	β	R^2	N	β	R^2	N	β	R^2	N
1760	-1.32^{***}	0.98	1,622	-1.37^{***}	1.00	1,459	-1.42^{***}	0.99	100
1860	-1.07^{***}	0.99	979	-1.10^{***}	1.00	881	-1.21^{***}	0.99	100

- Lower slope in 1860 (close to Zipf's Law), more uneven distribution.
- Still, Paris now on Zipf's prediction, other largest cities smaller than Zipf's prediction.

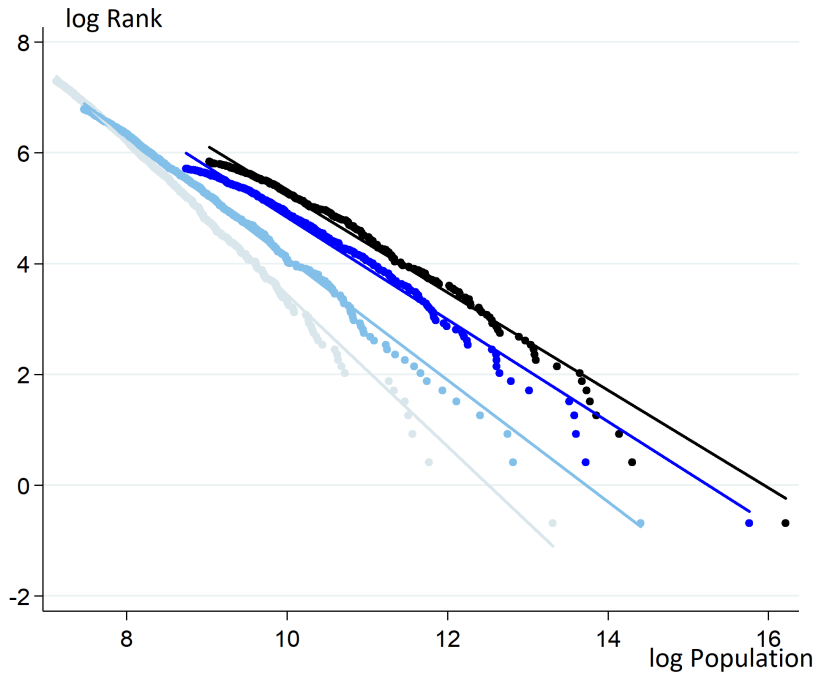
City size distribution: 1760, 1860, and 1960



	All cities			10% smallest excluded			100 Largest		
Year	β	R^2	N	β	R^2	N	β	R^2	N
1760	-1.32^{***}	0.98	1,622	-1.37^{***}	1.00	1,459	-1.42^{***}	0.99	100
1860	-1.07^{***}	0.99	979	-1.10^{***}	1.00	881	-1.21^{***}	0.99	100
1960	-0.86^{***}	0.96	337	-0.92^{***}	0.98	303	-1.07^{***}	0.99	100

- Even lower slope in 1960, more uneven distribution: Concentration in fewer larger cities.
- But Paris and other largest cities slightly below Zipf's prediction.

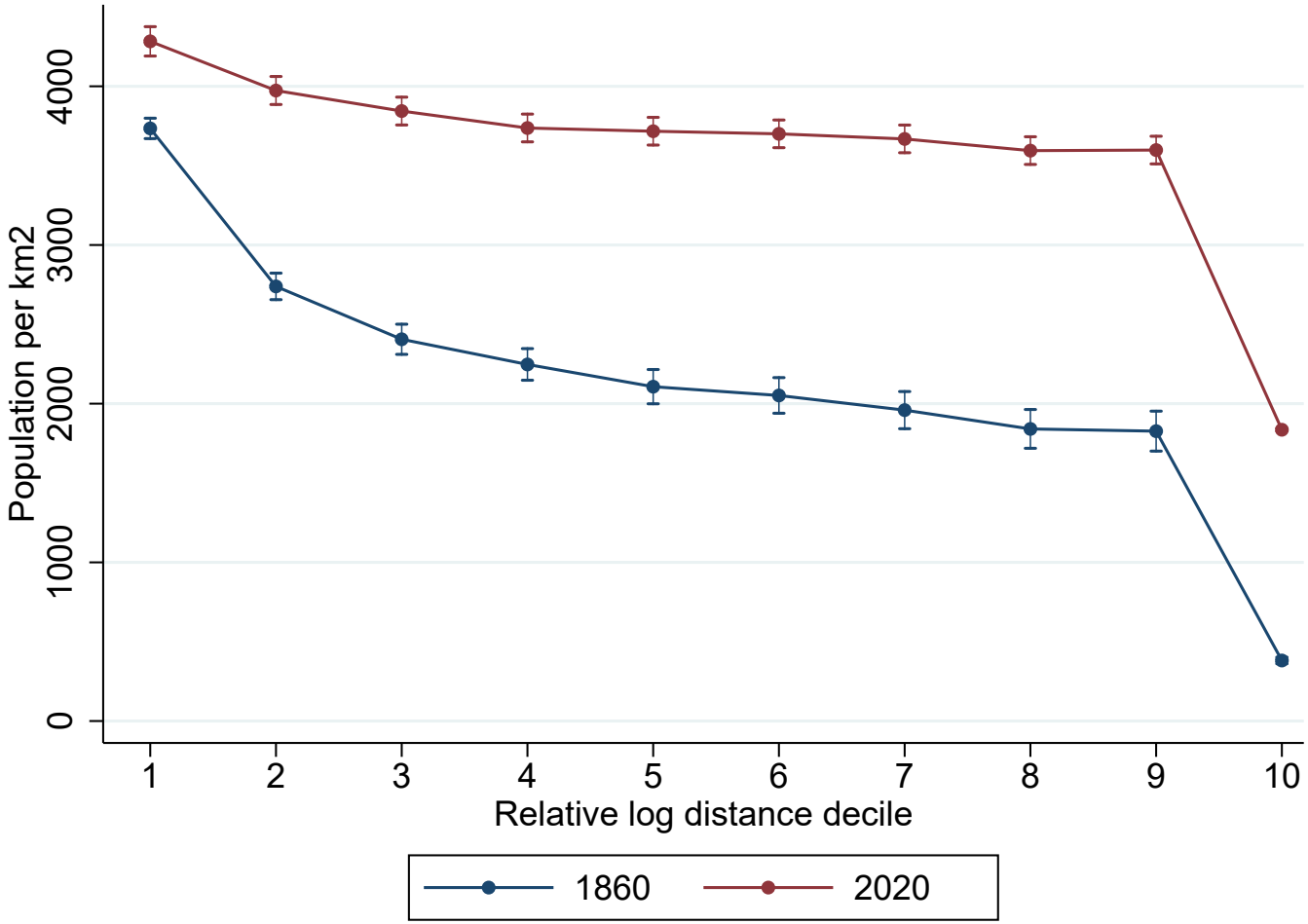
City size distribution: 1760, 1860, 1960 and 2020



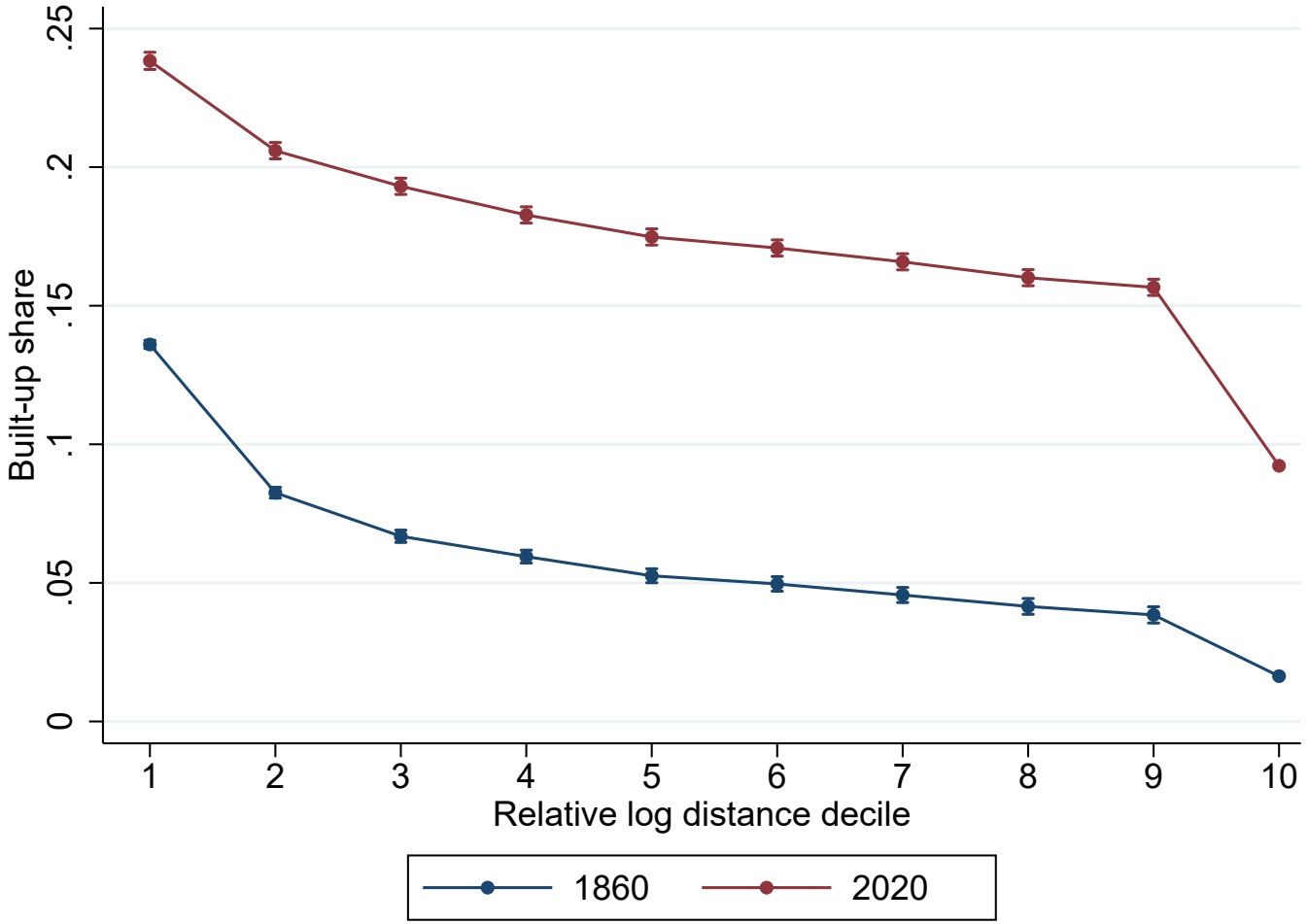
	All cities			10% smallest excluded			100 Largest		
Year	β	R^2	N	β	R^2	N	β	R^2	N
1760	-1.32^{***}	0.98	1,622	-1.37^{***}	1.00	1,459	-1.42^{***}	0.99	100
1860	-1.07^{***}	0.99	979	-1.10^{***}	1.00	881	-1.21^{***}	0.99	100
1960	-0.86^{***}	0.96	337	-0.92^{***}	0.98	303	-1.07^{***}	0.99	100
2020	-0.82^{***}	0.97	382	-0.88^{***}	0.98	343	-1.02^{***}	0.99	100

- Even lower slope in 2020, more uneven distribution: Concentration in fewer larger cities.
- But Paris and other largest cities slightly below Zipf's prediction.

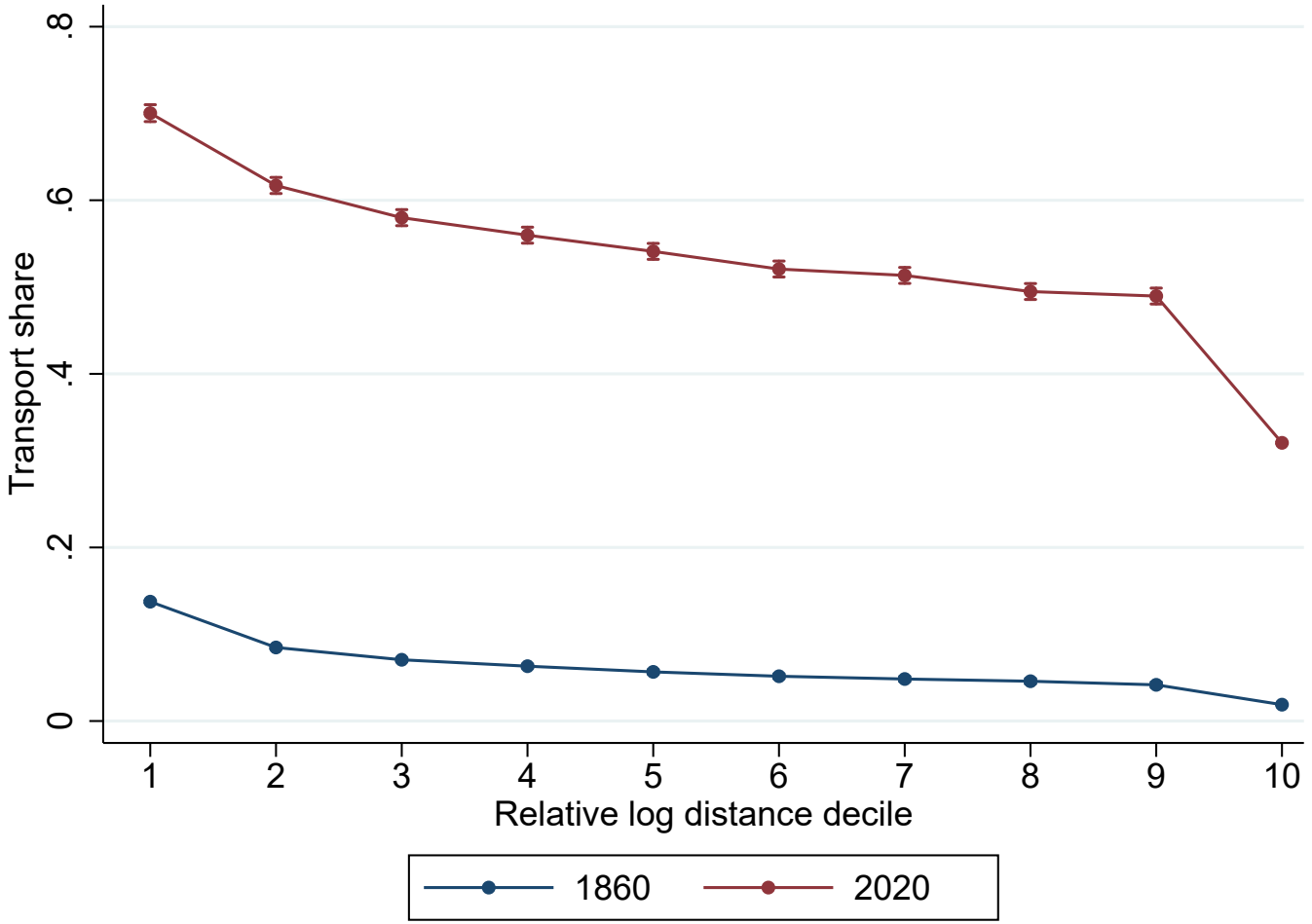
Population within-city gradient



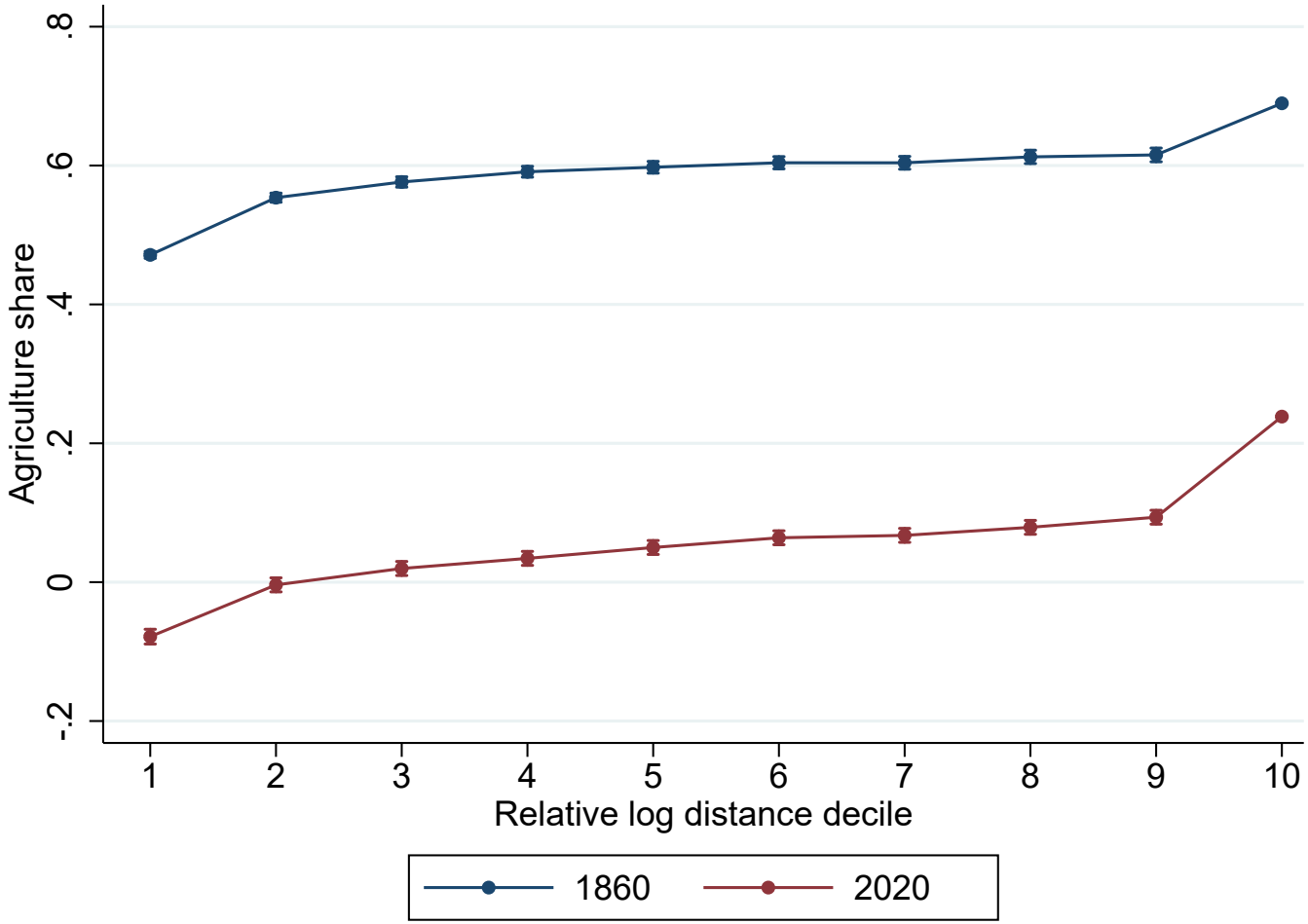
Built-up within-city gradient



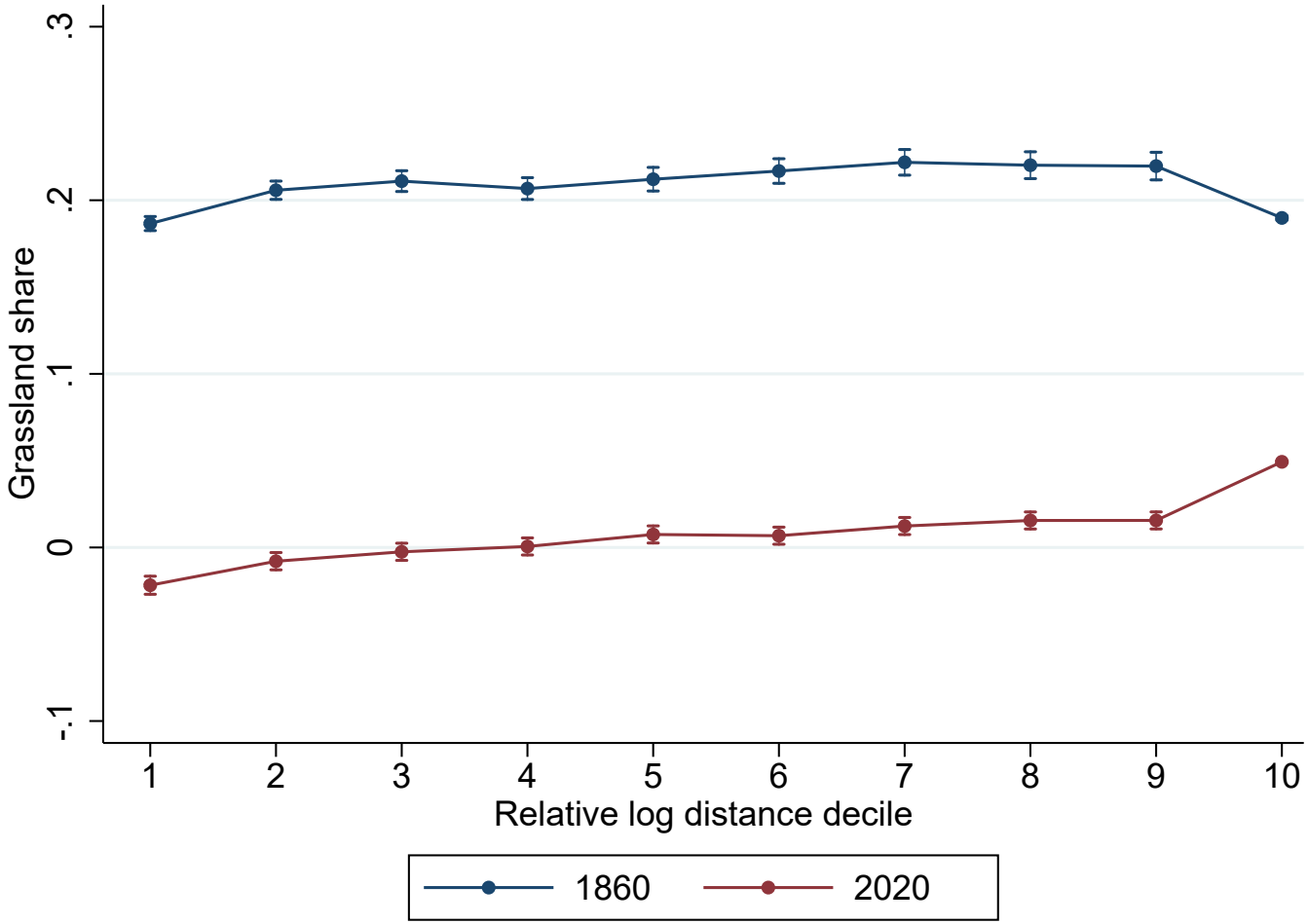
Streets, roads and railways within-city gradient



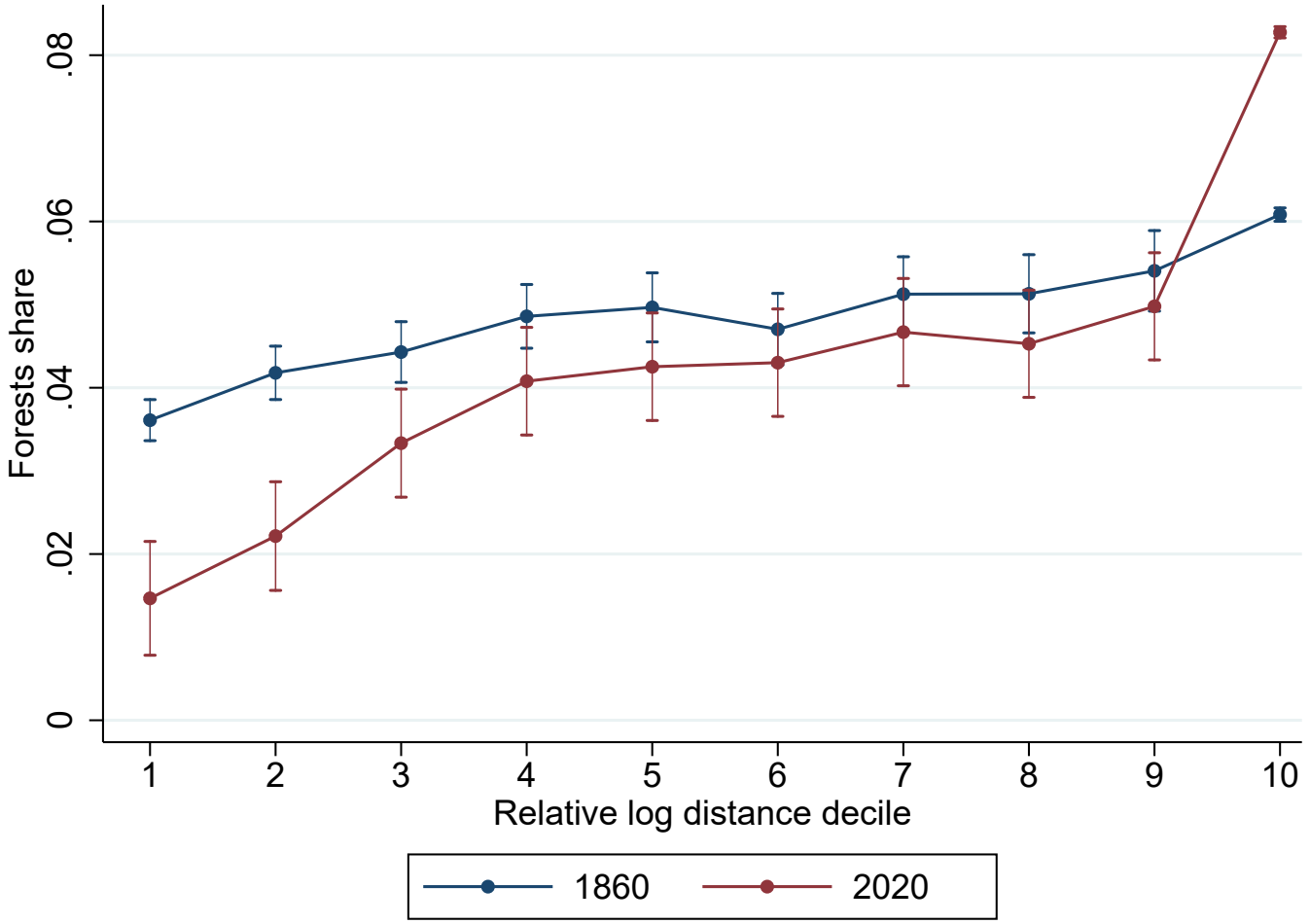
Agriculture within-city gradient



Pastures, meadows within-city gradient



Forests within-city gradient



Conclusion

- Machine learning strategy that can be adapted to all sort of maps and many exist for many countries (even if for more recent periods).
- Many interesting facts
  - 'Churning': Persistence of some cities but also disappearance and emergence.
  - Strong urbanisation with fewer and larger cities.
  - Large urban footprint increase with some density flattening within cities.
  - Monocentric then multi-centric market access structure moving South (East).
  - Cities took land from agriculture but most of the decline in crop land went to forests and pastures.

$\Rightarrow$ All consistent with standard urban economics predictions:

- Persistence due to non-reversible investments (infrastructure, built-up).
- Decline in commuting and trade costs

$\Rightarrow$ Third project: Estimation of a structural change model exploring the effects of agricultural productivity on urbanisation.