Eurostat data analysis

Geospatial data management project
A.Y. 2024/25

Federico Garegnani
Mat. 41712A

August 29, 2025

CONTENTS

1 INTRODUCTION

Access to healthcare services is a fundamental element in ensuring the right to health and reducing territorial inequalities. Everywhere, the distribution of the population and the location of healthcare facilities significantly affect how easily citizens can obtain medical care. Geographic distance, population density, and the availability of transport infrastructure play a crucial role in determining actual accessibility to essential services, thereby contributing to differences between urban, rural, and mountainous areas.

This project aims to analyze the ease of access to healthcare in Italy by integrating three datasets provided by Eurostat and the GISCO portal:

- *EU Population (Census 2021)*: contains information on the distribution of the European population in 1 km^2 grids, useful for assessing demographic density and spatial distribution.
- *Geographic Accessibility to Healthcare Services*: provides spatial data on distance and travel time to healthcare facilities, allowing the estimation of geographic accessibility to medical services.
- *EU Communes (Municipalities)*: includes the administrative boundaries of European municipalities, essential for aggregating and comparing data at the local level.

By combining and analyzing these sources, it will be possible to build a detailed picture of the Italian situation, highlighting both the areas with the best access to healthcare and those facing greater challenges.

Starting from these data, the study wants to answer the following questions:

1. Calculate the distribution of population in the country.
2. Determine the mean accessibility value to healthcare services in each municipality.
3. Determine the amount of population living in the municipalities where the mean accessibility level is low.
4. Compare the mean accessibility level value in 2020 and 2023, so as to determine whether there has been an improvement or not.

This report will be structured as follows:

- *Section 2*: description of used datasets.
- *Section 3*: data preprocessing.
- *Section 4*: demographic analysis.

2 DATASET PRESENTATION

For this project, three datasets provided by GISCO have been used. GISCO, the *Geographic Information System of the European Commission*, is the platform responsible for managing and disseminating geographical and cartographic data related to European statistics. It is maintained by Eurostat, the statistical office of the European Union.

The three dataset chosen for this study are the following:

- *Census grid 2021*
- *Geographic accessibility*
- *Communes*

2.1 *Census grid 2021*

The Census Grid 2021 dataset, published by Eurostat, provides a harmonized representation of the resident population across the European Union based on the reference year 2021. The dataset distributes population counts into a regular grid of 1 km² cells, offering a highly detailed spatial perspective on demographic distribution. Unlike administrative boundaries, which vary in size and shape across countries, the grid system ensures comparability and uniformity, making it particularly suitable for spatial analysis and cross-country studies.

Every cell contains the following demographic aggregated information:

- Distribution by sex (male, female)
- Age group (under 15 years, 15-64, 65 years and over)
- Current activity status (number of employed persons)
- Place of birth (place of birth in reporting country, place of birth in other EU country, place of birth elsewhere)
- Place of usual residence (total population)
- Place of usual residence one year prior to the census (usual residence unchanged, move within the reporting country, move from outside the reporting country)

Dataset uses ETRS89 as a reference system.

2.2 *Geographic accessibility*

The Geographic Accessibility to Healthcare Services dataset provides harmonized information on the ease of reaching healthcare facilities across the European Union and EFTA countries. The dataset is based on a grid of cells of 1 km² and contains indicators of accessibility expressed in terms of travel distance and travel time from each grid cell to the nearest healthcare service. Travel time data rely on transport time measures using road network.

This dataset provides two indicators for accessibility:

- driving time from each populated 1km² cell to the nearest healthcare service
- the average driving time to the three nearest services.

Each time indicator is available for the reference years 2020 and 2023.

2.3 *Communes*

The Communes dataset contains the official administrative boundaries of municipalities (also referred to as communes or local administrative units, LAU level 2) across the European Union. Municipalities represent the lowest level of administrative governance in most European countries, making this dataset an essential reference for analyses that require a fine territorial granularity.

Here is the list of available fields:

- COMM_ID: identification code of the municipality;
- CNTR_ID: here is simply the copy of CNTR_CODE, but it can also be an internal identification number of Eurostat;
- CNTR_CODE: country code as for ISO 3166-1 alpha-2;
- COMM_NAME: name of the municipality;
- NAME_ASCI: name of the municipality without special characters (ASCII characters);
- TRUE_FLAG: a technical attribute used to distinguish real polygons from those that exist only for cartographic or topological reasons;
- NSI_CODE: municipality code according to the National Statistics Institute;
- NAME_NSI: name of the municipality according to the National Statistics Institute;
- NAME_LATN: name of the municipality written with Latin characters. This is for countries which does not use Latin alphabet (Greece, Bulgaria, etc.);
- NUTS_CODE: NUTS CODE (Nomenclature of territorial units for statistics).

Dataset uses ETRS89 as a reference system.

3 DATA PREPROCESSING

Prior to the analysis, it was necessary to carry out several preprocessing steps. Since the three sources differ in structure and format, it was essential to harmonize them through a series of operations, such as data cleaning and integration of spatial and statistical attributes. Such preparatory work was crucial to enable the joint use of the datasets and to derive meaningful indicators, ultimately allowing us to address the research questions formulated at the beginning of this project.

3.1 *Filtering the datasets and creation of indexes*

We are going to analyse Italian situation and, for this reason, the first operation to perform is filtering the communes dataset so as to obtain a smaller dataset with only Italian municipalities.

The following query creates a new table called `it_communes`, obtained filtering the main table where the attribute `CNTR_CODE` assumes the value `IT`.

```
CREATE TABLE it_communes AS (
  SELECT *
  FROM "all_communes" c
  WHERE "CNTR_CODE" = 'IT'
);
```

An index of type R-Tree was created on the new table `it_communes`. The same type of index was also created on the other tables: `census_2021` and `grid_accessibility_health`.

```
CREATE INDEX it_communes_idx
ON it_communes USING GIST (geom);

CREATE INDEX census_idx
ON census_2021 USING GIST(geom);

CREATE INDEX accessibility_idx
ON grid_accessibility_health USING GIST(geom);
```

3.2 *Population calculation by municipality*

Since the dataset related to communes only contains borders information but no data regarding the population, it was necessary to integrate it with `census_2021` dataset. However, as described in the dataset presentation, they divide territory in different way: the `communes` dataset has the precise polygon representing commune border, while `census` dataset rely on a grid of cells. This obliged us to perform some form of approximation in order to compute commune population.

The problem is limited to the cells which intercept a commune border; here the municipal territory can overlap only partly with the cell area. There are two possible approximations, we can consider all the cells that intersect the municipal territory or we can consider only the cells entirely within the border. In the first case we will have an over-estimate of the population, since we are including people who actually live in another municipality, in the second case we have an under-estimation, since we are excluding some people. In both cases the estimate is too rough.

In order to compute the total population in the most possible precise way the following formula was used.

$$P_C = P_G \sum \frac{A(C \cap G)}{A(G)} \tag{1}$$

Here C represents the commune, G represents the cell of the grid and P the population.
This formula considers the fraction of cell area covered by the commune and multiply it for the total population living inside the cell. Obviously, even this formula finds a compromise, assuming people to be uniformly distributed on the cell area.

We are going to store the computed value for population in a new column of the table `it_communes` called `pop`. Here is the query to add the new column and insert the value for population for each commune in Italy. Since we know cells to have an area of 1 km², we can omit `ST_AREA(g.geom)` from the denominator, substituting it with $1'000'000$, that is 1 km² converted to m².

```
ALTER TABLE it_communes ADD COLUMN pop INTEGER;

WITH pop_calc AS (
  SELECT
    c."COMM_ID",
    ROUND(SUM(ST_Area(ST_Intersection(g.geom, c.geom)) /
     ↪  1000000 * g."T")) AS pop
  FROM census_2021 g JOIN it_communes c ON
   ↪  ST_Intersects(g.geom, c.geom)
  GROUP BY c."COMM_ID"
)
UPDATE it_communes c
SET pop = p.pop
FROM pop_calc p
WHERE c."COMM_ID" = p."COMM_ID";
```

3.3 *Mean accessibility to healthcare services*

To compute the mean accessibility value for each municipality we need to join data from `communes` and `grid_health_accessibility` datasets. Join operation is performed considering grid cells which intersect the commune area, data are then grouped by commune `ID` and the average of each column is finally computed. Here above the query used to define four more numeric columns and assign the mean value for health services accessibility. The meaning of each indicator (column name) is described in the dataset presentation (2.2).

```
ALTER table it_communes
ADD COLUMN health_2020_n1 REAL,
```

```
ADD COLUMN health_2020_n3 REAL,
ADD COLUMN health_2023_n1 REAL,
ADD COLUMN health_2023_n3 REAL;

WITH avg_t AS (
  SELECT
    c."COMM_ID",
    AVG(g.health_2020_n1) AS mean_health_2020_n1,
    AVG(g.health_2020_n3) AS mean_health_2020_n3,
    AVG(g.health_2023_n1) AS mean_health_2023_n1,
    AVG(g.health_2023_n3) AS mean_health_2023_n3
  FROM it_communes c JOIN grid_accessibility_health g ON
   ↪  st_intersects(c.geom, g.geom)
  GROUP BY c."COMM_ID"
)
UPDATE it_communes as c
SET
  mean_health_2020_n1 =
   ↪  ROUND(t.mean_health_2020_n1::NUMERIC, 2),
  mean_health_2020_n3 =
   ↪  ROUND(t.mean_health_2020_n3::NUMERIC, 2),
  mean_health_2023_n1 =
   ↪  ROUND(t.mean_health_2023_n1::NUMERIC, 2),
  mean_health_2023_n3 =
   ↪  ROUND(t.mean_health_2023_n3::NUMERIC, 2)
FROM avg_t t
WHERE t."COMM_ID"=c."COMM_ID";
```

4 STUDY OF THE ITALIAN TERRITORY

This section is a demographic analysis of Italian situation. We want to analyse how population is distributed throughout the country to later show how this aspect affects accessibility to healthcare services.

Figure 1 shows how population is distributed on a regional level, classifying each region with a graduated scale based on the total inhabitants of the region. Five classes have been defined, assigning each class the same interval. [1]

We can see from the figure that Lombardia is largely the most populated region, while Valle d'Aosta, Molise and Basilicata are the least populated.

1 Sturges's rule was applied to determine the appropriate number of bins for categorizing the data. It states

$$k = \lfloor log_2 \rfloor n + 1$$

where k represents the number of bins to be defined and n the total number of elements.
To decide how many classes to define for categorizing regions: $k = \lfloor \log_2 21 \rfloor + 1 = 5$.
Although Italy has 20 regions, the dataset counts 21 elements because Trentino-Alto Adige does not have a single NUTS-2 code. Instead, it is represented by two separate codes corresponding to its autonomous provinces.

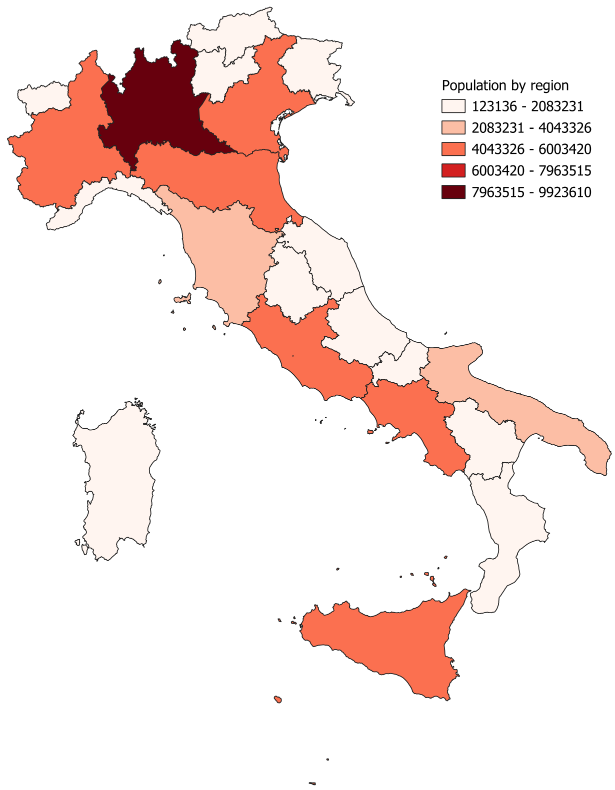


Figure 1: Italian population by region

By increasing the resolution we can analyse how the population is distributed at the municipal level. The graph on Figure 2 is very right-skewed, meaning that the great majority of Italian population lives in little municipalities, smaller than 50,000 inhabitants. The graph is so right-skewed we can not even see other columns.

The chart on Figure 3 illustrates the distribution of Italian municipalities by population size. On the left axis, the blue bars represent the number of municipalities within each population class. The right axis shows two lines: the orange line indicates the total population (in millions) within each class, while the red dashed line shows the cumulative population across classes.

The data reveal that the vast majority of Italian municipalities are small, with populations under 10,000 inhabitants. However, these small municipalities account for only a limited share of the total Italian population. In contrast, a relatively small number of large municipalities, those with more than 100,000 inhabitants, concentrate a significant proportion of the population. The cumulative line clearly highlights this imbalance: only a few classes at the upper end capture the majority of residents, despite representing a minority of municipalities overall.

Table 1 reports in numbers the situation described until now.

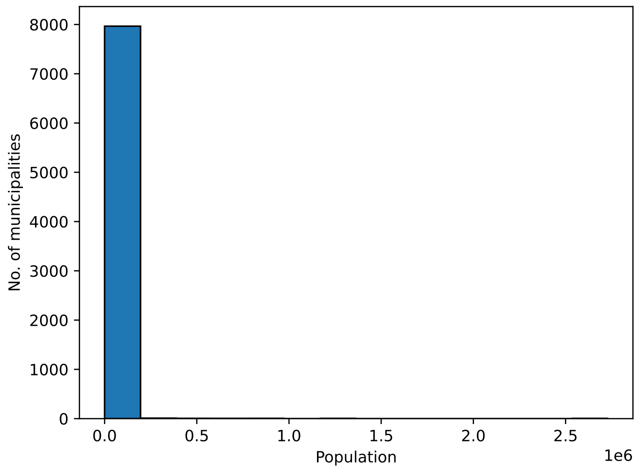


Figure 2: Distribution of population by municipality

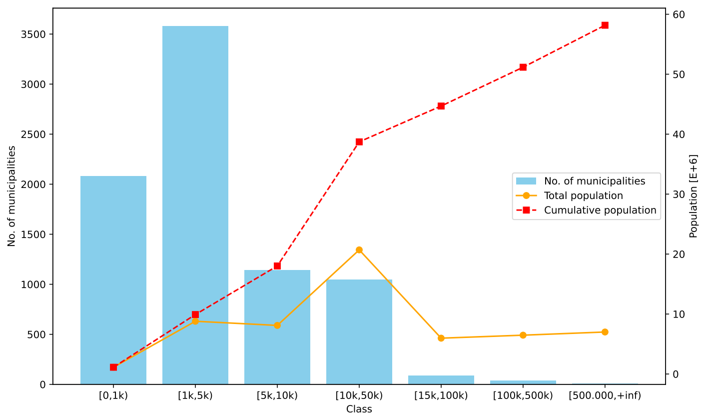


Figure 3: Distribution of Italian communes by population and total population living in communes of that class

This pattern emphasizes Italy's demographic structure, characterized by a high number of small towns and villages, alongside a concentration of population in a limited set of large urban centres.

5 HEALTHCARE SERVICES ACCESSIBILITY

5.1 *Italian overview*

Accessibility to healthcare services is a key determinant of population well-being and social equity. Ensuring that people can reach medical facilities within a reasonable time is essential for timely diagnosis, treatment, and prevention of health conditions. Differences in accessibility often reflect broader regional disparities, influencing quality of life, economic opportunities, and demographic sustainability. Studying healthcare accessibility therefore provides valuable insights for politician, helping to identify underserved areas and design more inclusive health systems.

Class	No. communes	Total pop.	% on total
< 1.000	2081	1 120 526	1.93
1.000 – 4.999	3580	8 795 876	15.12
5.000 – 9.999	1142	8 101 451	13.93
10.000 – 49.999	1048	20 700 471	35.59
50.000 – 99.999	88	5 973 913	10.27
100.000 – 499.999	38	6 472 224	11.13
>= 500.000	6	7 001 822	12.04

Table 1: Distribution of Italian communes by population.
For each category: number of municipalities falling within that class, total population residing in those municipalities, percentage of the national population contained in each class

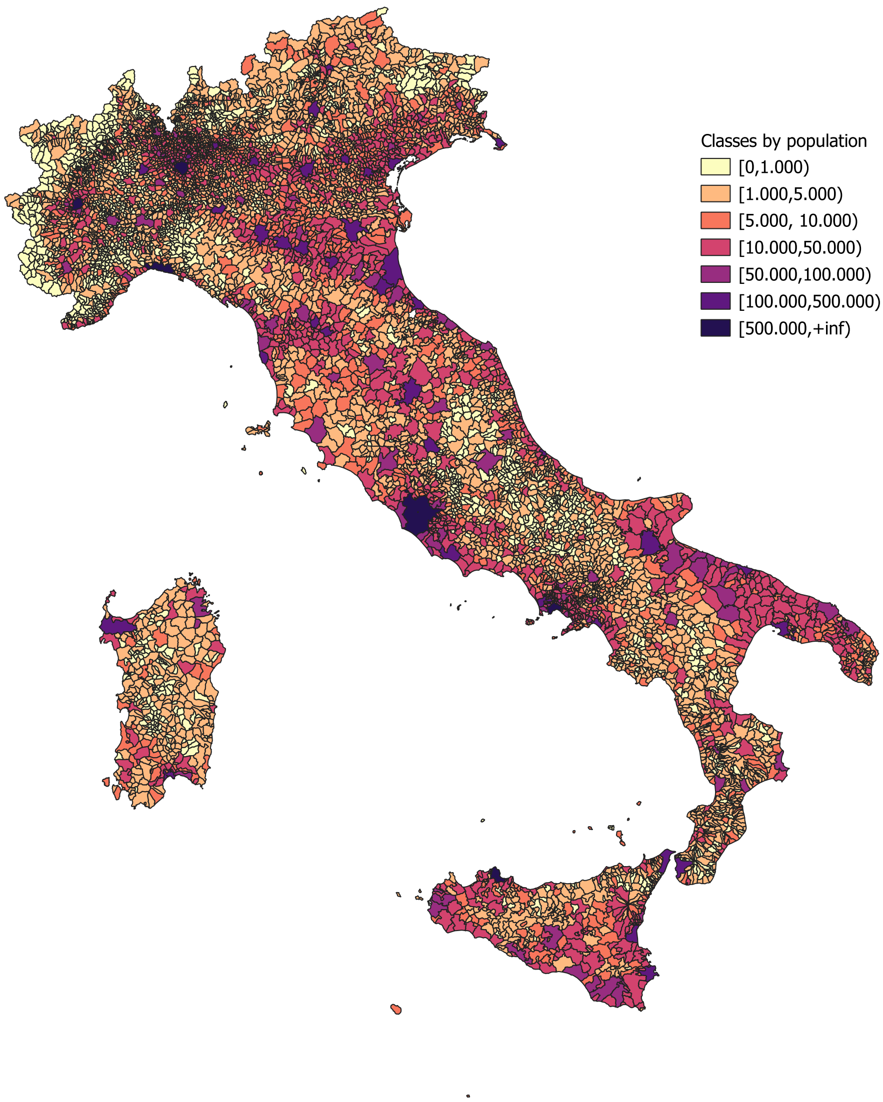


Figure 4: Classification of communes by population

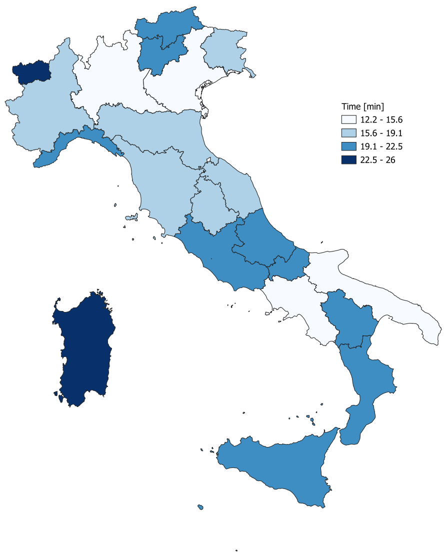


Figure 5: Accessibility to the nearest healthcare service by region in 2023

Figure 5 show the mean accessibility time to reach the nearest hospital structure for Italian regions, it is expressed as driving minutes to the nearest medical centre. Classification was realized categorizing mean times in intervals of equal dimension, technique which allows us to highlight regions that have better or worse average times. As can be seen from the graph, Northern Italy tends to have lighter colours, a sign of easier access to health centres. The most virtuous region of the peninsula proves to be Lombardia, with an average time of 12 minutes; on the other side of the ranking, Valle d'Aosta and Sardegna show the worst values, with average times up to 26 minutes.

If we look at the map showing the travel time to the three closest healthcare facilities (Figure 6), the situation changes slightly. For some regions, such as Trentino-Alto Adige and Basilicata, the situation worsens, placing them in the lowest class; for other regions, especially in central Italy, the situation improves slightly.

When reading these maps, one must not forget the nature of the Italian territory, which is mountainous in many areas. While urban and lowland areas often benefit from dense networks of medical facilities and efficient transport connections, many mountainous and rural regions face significant challenges. Steep terrain, dispersed settlements, and limited infrastructure can increase travel times to healthcare centres, creating disparities in access to essential services.

5.2 *North of Italy*

Let us now turn to northern Italy, a region chosen for its wide range of environments: an extensive plain, mountain ranges, and coastal areas. Access times have been organised into 4 categories:

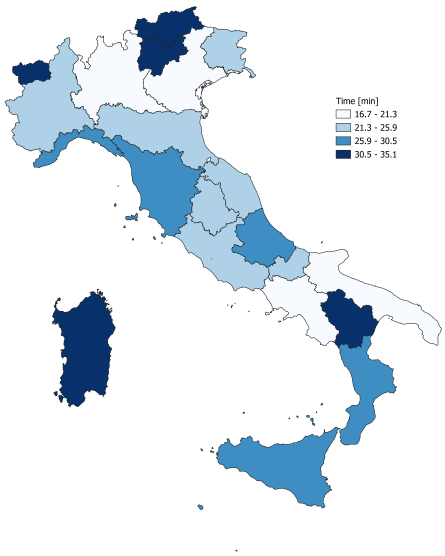


Figure 6: Accessibility to the 3 nearest healthcare services by region in 2023

No.	Travel time [min]	Accessibility
1	< 15	High
2	$[15, 30)$	Medium-high
3	$[30, 45)$	Medium-low
4	> 45	Low

From Figure 7 we note that Pianura Padana benefit of an excellent coverage in terms of health facilities, indeed it is the most populous and productive area in Italy, where there are no natural obstacles hampering connections or construction of new facilities.

This situation stays almost the same when considering accessibility to the three closest structures (Figure 8). The most densely populated areas do not experience any changes in category when considering the closest centre or the first three. This is the case of the provinces of Milan, Como, Bergamo, and the areas along A1 highway. People living in the aforementioned areas can reach a great number of medical centres driving less than 15 minutes.

The situation changes considerably when we look at the outer reaches of northern Italy, near the border, where the Italian territory is crossed by the Alpine mountain range, a natural border with neighbouring France, Switzerland, Austria, and Slovenia. The morphology of these areas, coupled with their lower population density and a less efficient road infrastructure, makes them much less accessible from a healthcare perspective. In mountainous areas near the border, travel times often exceed 15 minutes to the nearest healthcare facility, and consistently exceed 30 minutes to reach the three closest facilities.

As shown in the histogram in Figure 9, Northern Italy displays a favorable distribution of healthcare facilities, with 60% of municipalities located within a 15-minute drive of a facility and 92% within 30 minutes.

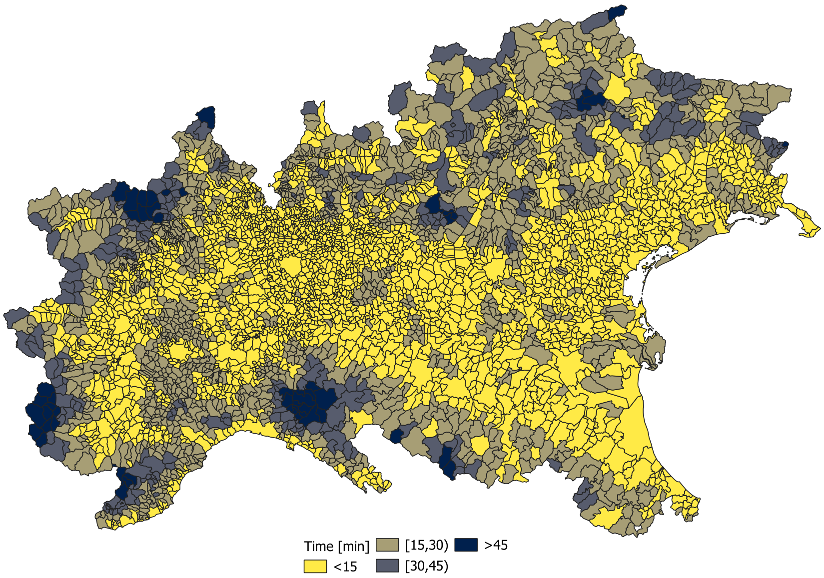


Figure 7: Accessibility by class to reach the closest healthcare service in 2023

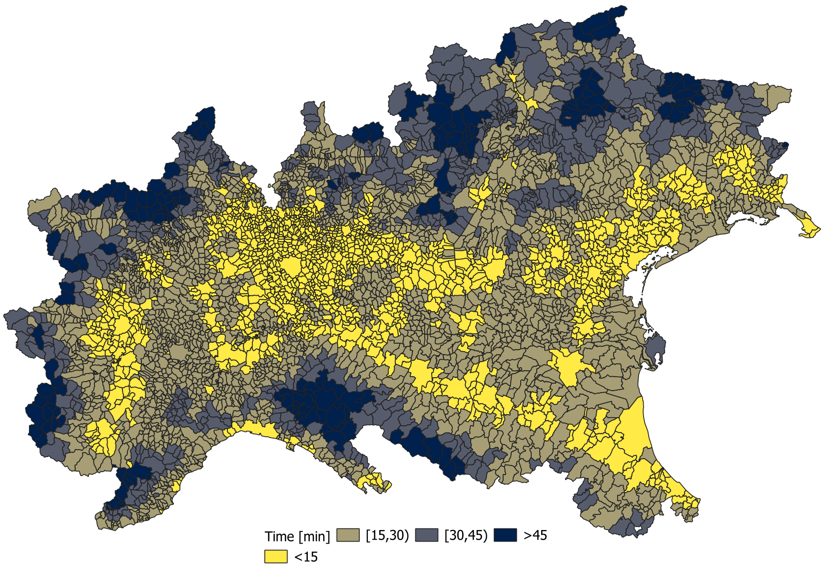


Figure 8: Accessibility by class to reach the 3 closest healthcare services in 2023

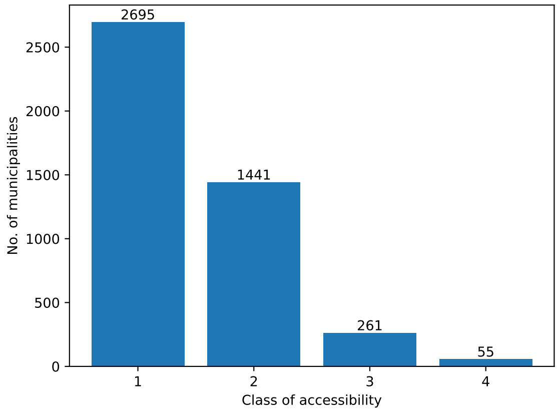


Figure 9: Distribution of accessibility classes in Northern Italy municipalities. Time to reach the closest healthcare service in 2023

5.3 *Population living in areas with poor healthcare accessibility*

We can calculate the portion of population living in the municipalities of each class, results are shown in the following table.

Class	Population	% of total
1	18 836 153	69.32
2	7 399 725	27.23
3	792 058	2.91
4	143 979	0.53
	27 171 915	

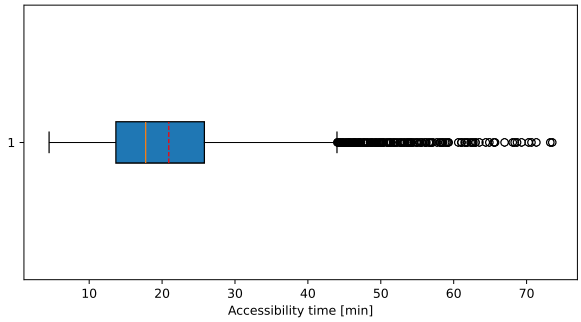
Based on time needed to reach the 3 closest health facilities, we consider those in classes 3 and 4 as municipalities with poor health accessibility. There are 936037 people living in communes of the two lower classes, equivalent to 3.44% of the population.

Overall, the distribution of the population across classes shows a generally favourable situation, with more than 96% of residents living in municipalities classified as 1 or 2, where access to healthcare facilities is relatively fast. Nevertheless, almost one million people (3.44% of the total population) live in municipalities belonging to classes 3 and 4, which indicates limited accessibility to healthcare services. As discussed before, these areas coincide with sparsely populated rural or mountain regions, where lower population density and geographical barriers constrain the availability of facilities.

5.4 *Change in healthcare accessibility between 2020 and 2023*

This section analyses the change in accessibility to reach the nearest healthcare facility between 2020 and 2023. The 7 intervals of change reported in Table 2 have been defined.

As can be seen from Figure 11, the majority of municipalities have not undergone significant changes in access times to medical facilities.



(a) Boxplot of accessibility times

Index	Value
Q1 (25%)	13.66
Median (50%)	17.75
Q3 (75%)	25.8
Mean	20.92
IQR	12.14

(b) Boxplot statistics for accessibility times

Figure 10: Distribution of municipal accessibility times to the three closest healthcare facilities

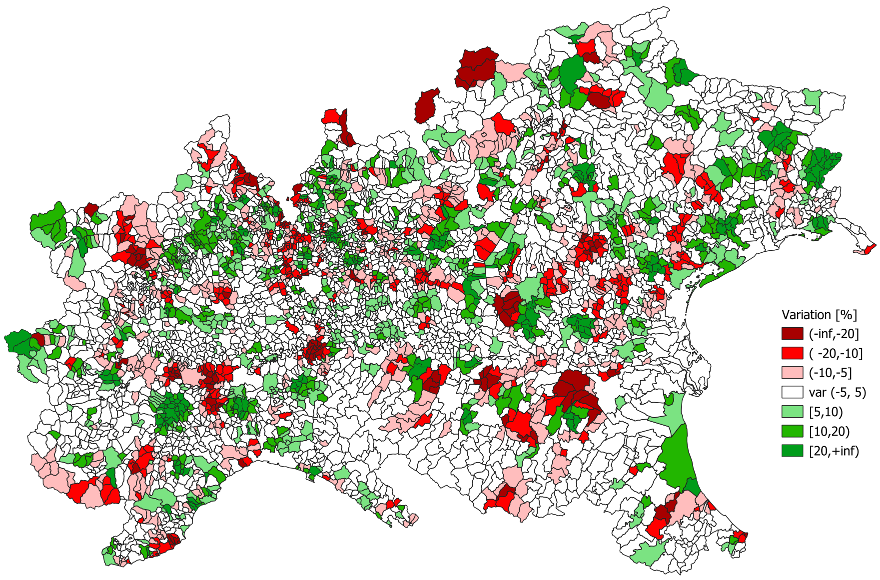


Figure 11: Percentage variation of accessibility times to reach the closest medical centre from 2020 to 2023 by commune

Class	Interval [%]	No. communes
Sharp worsening	$(-\infty; 20]$	135
Relevant worsening	$(-20; -10]$	348
Slight worsening	$(-10; -5]$	586
No change	$(-5; 5)$	2298
Slight improvement	$[5; 10)$	494
Relevant improvement	$[10; 20)$	355
Sharp improvement	$[20; +\infty)$	236

Table 2: Number of municipalities in each class of percentage change in access time to the nearest health centre

Percentage change in time was obtained with the following query, using this formula:

$$\Delta_{n3}[\%] = \frac{\bar{t}_{2023n1} - \bar{t}_{2020n1}}{\bar{t}_{2020n1}} * 100$$

```
SELECT
  c."COMM_ID",
  c."COMM_NAME",
  c.mean_health_2020_n1,
  c.mean_health_2023_n1,
  ROUND(((c.mean_health_2023_n1 - c.mean_health_2020_n1) /
  ↪  c.mean_health_2020_n1 * 100.0)::NUMERIC, 2) as
  ↪  variation_n1,
  c.geom
FROM it_communes c
WHERE SUBSTRING(c."NUTS_CODE" FROM 1 FOR 3) IN
↪  ('ITC','ITH');
```

We report here the count of the municipalities that fall into each class in Table 2.