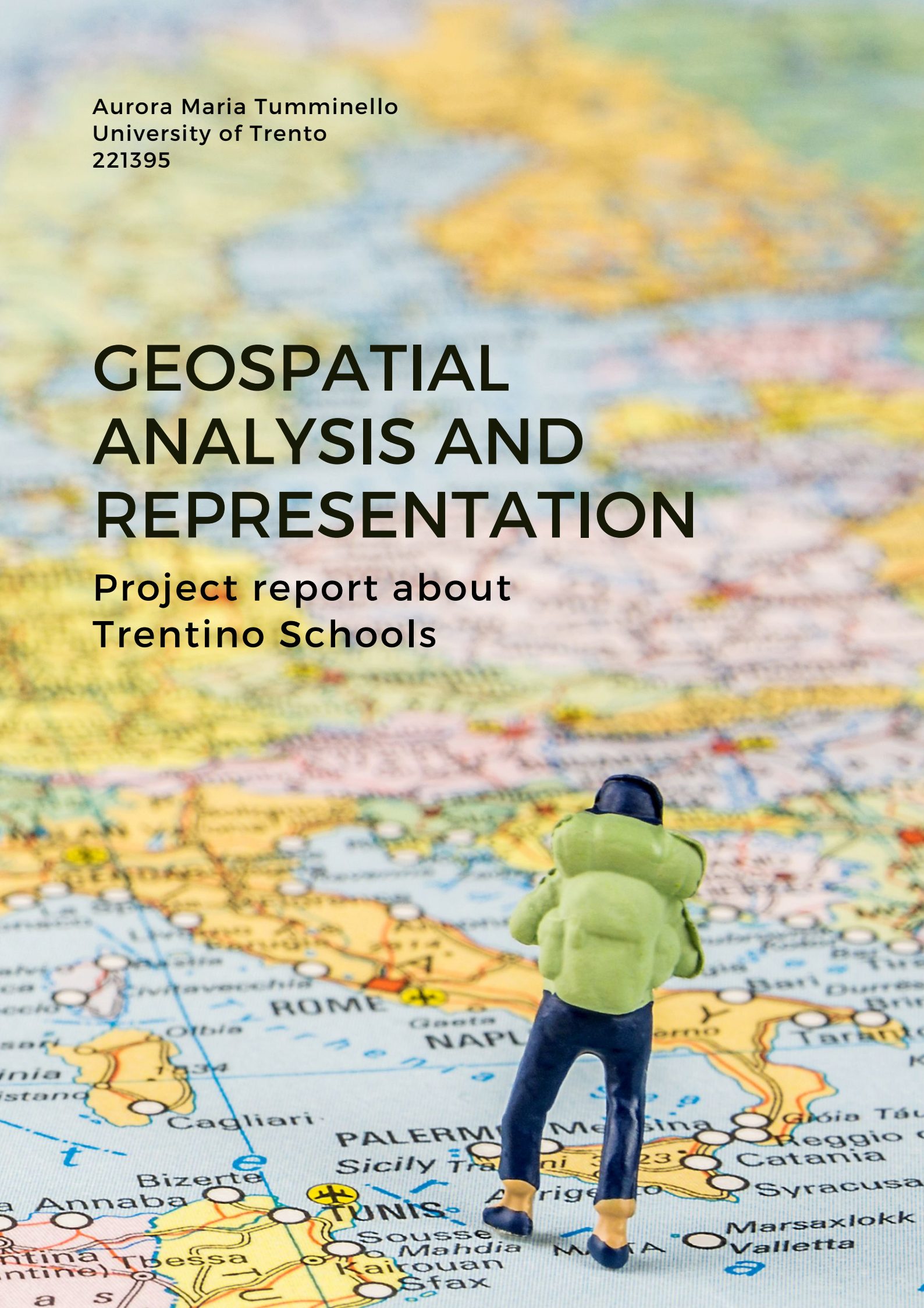
Aurora Maria Tumminello
University of Trento
221395

GEOSPATIAL ANALYSIS AND REPRESENTATION

Project report about Trentino Schools



Introduction

This report will guide us to the project of Geospatial Analysis and Representation for Data Science, which focuses on the school system in the province of Trento (also called Trentino). This report aims to make light of the steps that led to the creation of the website as a means to the storytelling situation about schools in the local territory.

⚠️ *Note that you can visualize the code and its output by clicking on the text "Notebook 0x on Github" near the Github icon. Some of the links will redirect you to Github, while others will open a nbviewer window to correctly visualize HTML outputs too (disabled on Github).*

Data Preparation

 **Notebook 01 on Github**

Initially, the idea was to use the data published on the ministerial of education website (MIUR), but the majority of data about Trentino were not available or not accurate enough as those found in vivoscuola, the official website for the school system in the province of Trento. It allows users to openly download data about schools, their addresses and contact information, from kindergartens to professional schools.

On the other hand, aprilascuola is a project in collaboration with vivoscuola, which offers an API to download not only schools data but also the number of students and classes per school, per grade and sometimes per academic year. Despite this dataset does not consider kindergartens, it contains provincial codes necessary for further steps to retrieve students' information. In the end, both these two sources (`data/aprilascuola.json` and `data/vivoscuola.csv`) were merged into `data/Trentino/schools/`, taking address information from vivoscuola (more accurate) and geographical coordinates from aprilascuola (when present).

Geocoding

 **Notebook 02 on Github**

Starting from the mere address, geocoding is a technique that allows the retrieval of geographical coordinates from a specific provider. In the second notebook, the dataset has been geocoded through 4 different steps:

- Considering the name, address, postal code and city, searching for matches using **nominatim** (OpenStreetMap provider). Since few schools have found such a specific match, additional steps are required;
- Considering address, postal code and city, searching for matches using **ArcGIS** provider. Most of the schools find their match, but still a lot of them have missing coords;
- Considering address, postal code and city, searching for matches through nominatim, again, excluding those points where the considered address is just the postal code or null (too generic or missing coordinates);

- 8 schools remain without a match. Although they have been manually inserted inside OpenStreetMap after a little online research, these didactic units are not found. That's why manual coordinates are provided.

After merging the original dataset with the ones obtained by geocoding schools addresses with missing location information, we get the final geojson file that will be used in further notebooks (`data/aggregated_data_per_municipality`).

Clustering Map

Notebook 03 on Github

Since we detain schools' data with their geographic locations, we can create a map that shows schools' markers with their information (e.g. name, institute, address, contact information etc). Considering that we are talking about over 700 schools, these data points may lead to an overwhelmed visualization of Trentino Province, leading to overlapping points and an unclear map. Instead, the idea was to create a **clustering map**, such that based on the level of zooming different markers can be clustered together into a bubble that counts how many points are aggregated. By zooming in, bubbles will get lighter and aggregate fewer points, until we reach a zoom level such that there is at most one school in the neighbourhood and its marker is visible.

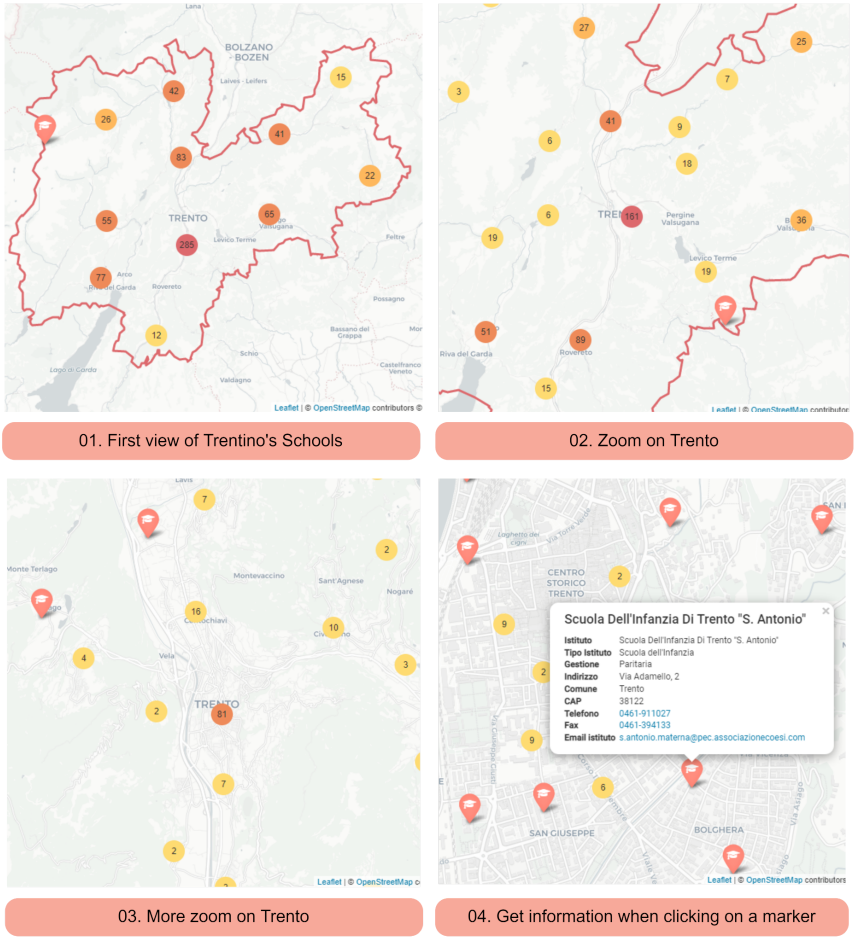


Figure 1: The clustering map

As an additional feature, the magnifying glass in the upper left corner allows us to search for a specific school based on its name. It immediately opens the related popup, showing the school's information. Notice that information such as phone numbers, fax, emails and websites are clickable and redirect to phone calls, emails and browser pages respectively.

Openstreetmap

Notebook 04 on Github

In previous steps, we preprocessed schools' data, got their position and plotted them on the map. The next step is to build a customized map for each school, containing some points of interest (POI) inside its neighbourhood (a radius of 1 km). We define as points of interest those belonging to one of these six categories: **Culture** (e.g. library, museums), **Sport** (e.g. gym, tennis court), **Food** (e.g. bar, restaurants, supermarkets), **Transport** (bus stops and train stations), **Outdoor** (green areas), **Healthcare** (e.g. hospitals, pharmacies) and **Utilities** (e.g. drinking water sources and copy shops). The points of interest, together with their type of amenity, opening hours and additional information, are provided by pyrosm, which allows us to communicate with OpenStreetMap. However, transport data were retrieved from Trenitalia and Trentino Trasporti.

Notice that the categorization of all points of interest is subjective, based on the type of amenities that were first retrieved (printed in the notebook as `list(set(pois['category']))`). Also, since the original category is in English, to build an Italian map, a JSON file with traductions was created (as `data/traduzioni.json`).

Initially, the idea was to create a single map with all POI and schools, but the loading of the map with all the markers in the Trentino Province required too much time (we're talking about nearly 100k markers). Therefore, a map for each school was created in `viz/pois`, with a yellow circle around the school indicating its surrounding neighbourhood (see Figure 2).

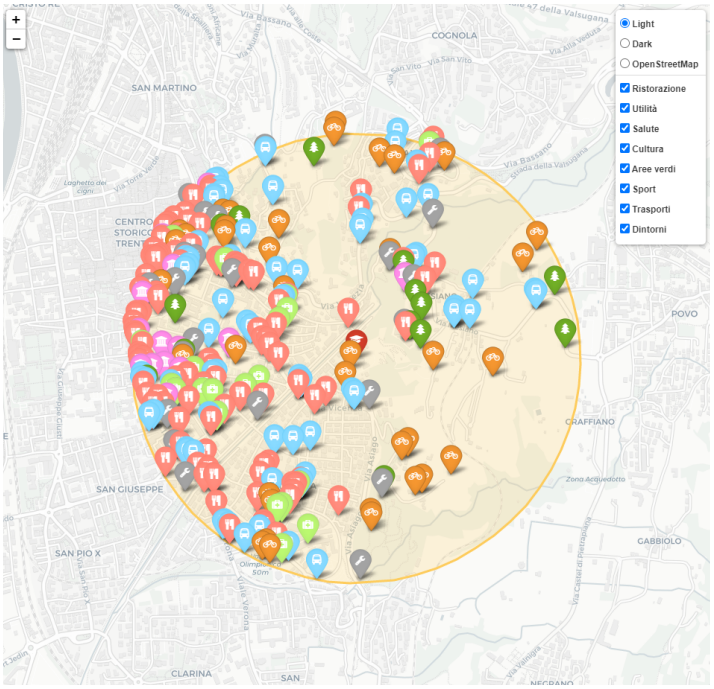


Figure 2: POI map around Liceo Scientifico G. Galilei in Trento

From the Layer Control in the upper right corner, it is possible to switch light/dark/osm modality, but also to (de)select the categories forehead mentioned. When clicking on a random marker, we obtain the name, opening hours, category, phone number, email or website (when present) and the distance from the school (both in terms of minutes and meters).

Isochrones

Notebook 05 on Github

The idea of this notebook is to explore the connection between a specific school and its neighbourhood in terms of streets and the amount of time necessary to reach them. For exploring such connections, **isochrone** maps will be used, which depict the area accessible from a point within a certain time threshold. In this case, there will be four different thresholds: 5, 10, 15 and 20 minutes, respectively coloured in yellow, orange, red and dark red.

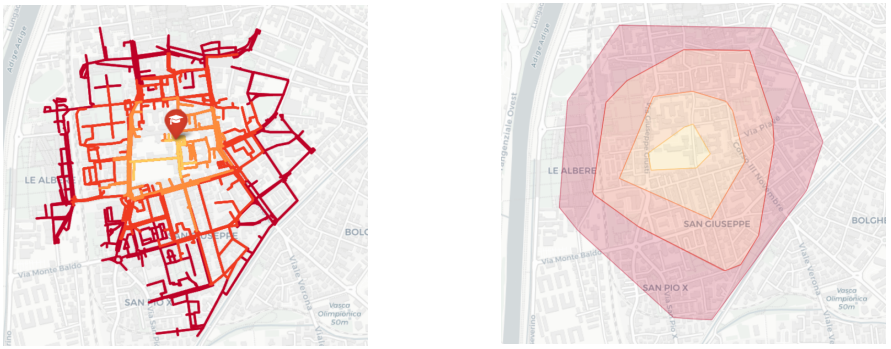


Figure 3: Isochrone maps in the route (left) and polygons (right) forms

Isochrones can be reproduced both as routes and polygons (as depicted in figure 3). On the website, route isochrones are considered, since more accurate than polygons, which instead show the reachable area using convex conjunction of the most external points. With route isochrones, the user can explore all reachable streets around the school.

Despite isochrones are easily buildable through **OSMnx** as static images, the notebook provides some functions to deploy interactive maps through folium, with some limitations, such that it is not possible to create layers because the subgraphs computed based on the time travel cannot be added as layers to a map, but only as of the map itself. Instead, the original idea was to create two select options inside the map: one for the means of transport (walk, bike or car) and the other for the time (5, 10, 15 and 20). The patch used to cover the limitations of folium joined to OSMnx was to generate an isochrone per school inside the `viz/isochrones/routes` folder, with all timings and to select the means of transport through the **website**.

Students

Notebook 06 on Github

In previous notebooks, we have worked with schools' data, without considering students. Notebook 06 focuses on students, on comparing them in terms of numbers with the total population, the population under 21 years old and the number of classes.

To do that, **ISTAT** population data (updated to 2019) of Trentino resulted necessary and divided based on gender, civil state and age. Then these data were merged with schools data and aggregated by the

municipality to get the total number of schools per area. To get students' data instead there are two options:

1. Scrape students and classes' numbers from Aprilascuola: given the specific id of the school (which is available only for schools inside the apri dataset), we can obtain the total number of students and classes, saved in `data/Trentino/schools/students.pkl`;
2. Since Aprilascuola API provides no instruction on how to use it, initially an alternative way to get data was to ask the department of education and culture of Trentino Province. After a few meetings, the director of such department, Francesco Pisanu, kindly provided the students' data updated to December 2021 (so fresher data than the ones published in aprilascuola at the start of the scholastic year). For this reason, these data were preferred to the scraped ones. Despite that, the code for scraping is still available. These data can be found at `data/population/students_per_school.csv`.

After aggregating the students' numbers with municipalities' and computing some statistics (e.g. ratio of students per population), an interactive choropleth map was built, with different layers: number of students, total population, number of schools, the ratio of students over the population and the ratio of students over the population under 21 years old.

Since each municipality belongs to a specific community, additional space was dedicated to their exploration. An initial map was created to make the user understand where communities are located and which municipalities they embody. Then, to observe these same data (i.e. students, schools, population, young population and classes) on communities, a treemap for each feature was built and saved in the `viz/trees` path.

Spatial Regression

Notebook 07 on Github

This last notebook focuses on the R part of the project, aimed at spatial and statistical analysis of the Trentino schools dataset. It has been saved in the `.Rmd` format, but also in the `.html` one, to visualize the output as if it was a notebook. We will use the ESRI shapefile saved at `data/aggregated_data_per_municipality` which contains, per each municipality, the number of schools, students, classes, population, young population, plus some ratios and means.

Initially, it was discovered that some municipalities are represented as Multipoly objects or as bizarre shapes and therefore their centroid (i.e. the representative point) may not be within their boundaries (such as *Tione di Trento* and *Luserna*). Then, different **definitions of the neighbourhood** were examined: K-nearest neighbourhood (**KNN**) to explore the k closest neighbours; Critical cut-off (**CCO**) to get the neighbours whose distance is equal to or less than the minimum necessary distance to make all municipalities connected to at least one another; Contiguity based approach (**CBA**), which considers as neighbours those spatial units that share a common boundary. These three definitions apply to municipalities, not to school points.

Based on every neighbourhood definition, the **Moran's I test for spatial autocorrelation** is computed on school-related features (e.g. Number of schools, mean of students per class). Among all the combinations of feature-neighbourhood, just three features obtain high values for Moran's I statistics: **mean students per school**, **population under 20 over the total population** and **students over the population under 20**.

Then we proceed with the **Moran's I test for spatial autocorrelation in OLS residuals**: while mean students per class results in a negative spatial autocorrelation with a high p-value (and therefore no significance), we obtain a Moran's statistic for the KNN neighbourhood with low p-value with the population under 20 over the total population, indicating a violation of the assumption of independence of residuals.

Whereas the Moran's I statistic is a global measure, the local analysis through **Moran's scatterplot** and **Local Moran's I** allow us to identify local patterns of spatial autocorrelation. Despite the scatterplots might provide an intuitive overview of local patterns of Trentino, Local Moran's I permit us to assess the significance of such patterns:

- *Population under 20 over Total Population*: **Sagron Mis**, **Cinte Tesino** and **Castello Tesino** have the highest Local Moran's I with the lowest p-value (mainly identified through KNN and CCO). All three municipalities are in the LL quadrant (i.e. low value for students over the population), while others, such as **Novaledo** and **Vignola-Falesina**, may not show such high significance but detain a high value for this ratio;
- *Mean students per school*: **Trento**, **Rovereto**, **Mori**, **Mezzocorona**, **Avio**, **Lavis** and **Ala** seem to be the municipalities with more students per school, while **Drena**, **Fornace**, **Ronzo-Chienis**, **Vallarsa** and **Terragnolo** are those municipalities that have few students but surrounded by municipalities with a lot of them. Most of the statistically significant municipalities are in the Valsugana or around the Rovereto/Riva del Garda zone;
- *Students over Population under 20*: Since some municipalities host more students than those who actually live in that area, one may think that students may need to move from their city to another to go to school every day. The most critical situations are around **Tione di Trento**, **Ossana**, **Cles** and **Rovereto**, which have a high number of students over its under 20 population, while **Valdaone**, **Selle Giudicarie**, **Porte di Rendena** and **Pelugo** lack students. Something similar but limited happens between Canazei and Giovanni di Fassa, since their closeness and the gap of students in the first against the abundance in the second may imply that students in Canazei move to Giovanni di Fassa to go to school. In this case, few municipalities show statistical significance for Moran's I.

The last part of the notebook is about **Spatial Regression Models**, to assess spatial spillovers (whenever a change in a spatial unit influences a change in another). Both local and global models were created, at first about the population under 20 over the total population and then on the mean of students per class:

- In the first case, a positive significant total impact on the proportion of the population under 20 over the total population has been found through the **mean number of students per school**, implying that higher levels of potential students in the territory lead to an increase in the mean students per school in the same neighbourhood.
- In the latter case, a positive significant impact on the **mean students per class** has been found through the number of students over the population under 20 and the number of schools, implying that higher levels of students per class in the territory lead to an increase in the number of schools and students ratio over young population in the same neighbourhood. The same applies to the number of classes, with a negative impact since the more classes there are, the more spread students will be among them.

Conclusion

Before actually starting this project, the focus was on the entire Italian School System, but, given the lack of quality in MIUR's data and the complexity of depicting the national scholastic situation, the final decision was to highlight the local territory of Trentino Province. This was revealed to be a potentially huge and interesting project, with many expansion possibilities. Since its implicit complexity, the website about Trentino Schools may serve to present the project in a more visually attractive form to users, with both a global and a local perspective on the territory. Many things could be improved, such as the isochrones in layers, the categorization of OpenStreetMap points of interest, additional features to the aggregated municipalities dataset, an investigation of disabled students and schools' accessibility etc. However, this project in its totality sets itself as a starting point for additional research about the topic and a possible expansion toward a national analysis of the school system.