Università degli studi di Milano-Bicocca

Data Science Lab for Smart Cities

Final Essay

Analysis of Urban Sustainability in Milan: GeoSpatial Machine Learning Developed on Air Pollution and Urban Environment

Authors:
Edoardo Fava - 851665 - e.fava4@campus.unimib.it
Davide Prati - 845926 - d.prati@campus.unimib.it

July 8, 2024

Contents

Abstract

The concept of smart cities embodies a dynamic and evolving approach to urban development, driven by goals such as environmental sustainability and enhanced quality of life. This paper explores Milan as a case study in urban transformation, emphasizing initiatives to address climate change, digital innovation, and sustainable mobility. Milan's transformation towards a "15-minute city", where essential services are easily accessible, highlights the importance of improving air quality and enhancing green spaces to create a healthier urban environment.
We will analyze air quality data from 12 monitoring stations across the city, sourced from Milan's open data and additional community sensors, to identify correlations between air quality and existing green spaces and buildings. We will try to determine, using ML techniques, the possible values of pollution across the city in places where currently there are no stations available. We will also determine how much the additional presence of a new hypothetical green area, as a source of air purification, can improve air quality and consequently the city's quality of life.
Attention will be given to the challenges of integrating green spaces in degraded areas to ensure they contribute positively to the urban environment. The goal is to promote a sustainable and livable city, addressing both environmental and social dimensions of urban development.

To do this, we will conduct data research from various sources, explore and prepare them, and augment them, all using the Python library GeoPandas [1] to work with geospatial data. To consider buildings and green areas' positions in the process, with respect to each station they are aggregated in areas considering both distance and angle.
After the process of data augmentation, we will have more than 400 covariates, undergoing an heavy feature selection process. Among various ML algorithms, the best one will be chosen on the basis of interpretability and R^2. At this point, based on real and realistic scenarios, we will estimate the air pollution levels that could be recorded by a station in the Darsena area, and the improvement in air quality that would result from the creation of a new park at Scalo Farini, as an example to show that our model can be used for such tasks of prediction and impact assessment.

1 Problem Description and Indicators

1.1 Smart Cities and sustainability

When discussing smart cities, both in literature and in common language, the term refers to a complex entity whose definition is often intuitive and not entirely standardized. This is because it tends to change rapidly over time and also depending on the socio-economic context. However, it is important to remember that one of the historical and cultural foundations of this idea is none other than one of the main environmental sustainability treaties in human history, namely the Kyoto Protocol of 2005 [2] [3]. Equally significant are documents like the Covenant of Mayors (2008) [4], where environmental sustainability is explicitly stated as the primary goal of a smart city.

Taking a step back, the concept of 'sustainable development' [5] is a wider concept that was introduced in the 80s and aims to promote a model of development that consists in a balance between social equality, economic growth and environmental preservation and proliferation.
Sustainability represents a holistic and long-term vision, which aims to match human activities and natural aspects in the most intelligent and advantageous way possible, in a way whereby man does not oppress the natural environment, but instead gains as many benefits as possible from it. This problem is clearly amplified in the context of 'megacities' [6], in which life is frenetic, physical spaces are reduced and the population density is such that it has required a major overbuilding operation over the years.

In this scenario, it becomes fundamental another concept, the one of urban sustainability [5]: a model of urban development that focuses on the realization of urban contexts designed with minimal environmental impact and liveable in a healthy way. To do so, it is necessary an efficient use of naturally available environmental resources and smart urban planification. Sustainable cities are cities that manage to maximize energetic efficience, reduce waste and pollution, support most of the forms of renewable energy, promote sustainable mobility and preserve urban ecosystems.
The key concept is that all this, in addition to helping the environment, produces an improvement in the lifestyle of citizens or inhabitants of the urban area, with consequent social and economic revaluation of the same.

1.2 Milan environmental situation and problematics

Delving into the details of our project, the city we will focus on is Milan. Milan is undergoing a profound transformation, driven by the need to address emerging challenges such as climate change, digital transformation, and sustainable mobility [6]. In the wake of the Covid-19 pandemic, the city administration, through the document "Milan 2020 Adaptation Strategy" [7], has launched a series of initiatives aimed at reorganizing the city's spaces and schedules, promoting the use of bicycles, walking, and the improvement of public spaces. This program aims to create a city where all residents can meet most of their needs with services located within a short distance from their homes. The goal is to rethink the organization of the city in terms of connectivity and urban fabric, promoting a more easily sustainable neighborhood life.
This idea can be summarized with the expression "15-minute city", where all essential services are reachable within a minimal travel time, thus reducing the need for long commutes.

A fundamental aspect of this transformation concerns the environment and quality of life, with particular attention to air quality. Green areas play a crucial role in improving air quality by reducing the presence of atmospheric pollutants and contributing to a healthier urban environment. The presence of parks, gardens, and green spaces is associated with a reduction in concentrations of pollutants such as particulates, as well as offering psychological and social benefits to the population.

It should also be considered that Milan is a city where, due to the morphological and geographical characteristics of its location, temperatures, especially in the summer, tend to be very high for many days of the year. This phenomenon is expected to intensify increasingly due to global warming, and in projection, it is expected that the average annual temperatures will increase by 4 degrees Celsius over 150 years, from 1901 to 2050 [8].

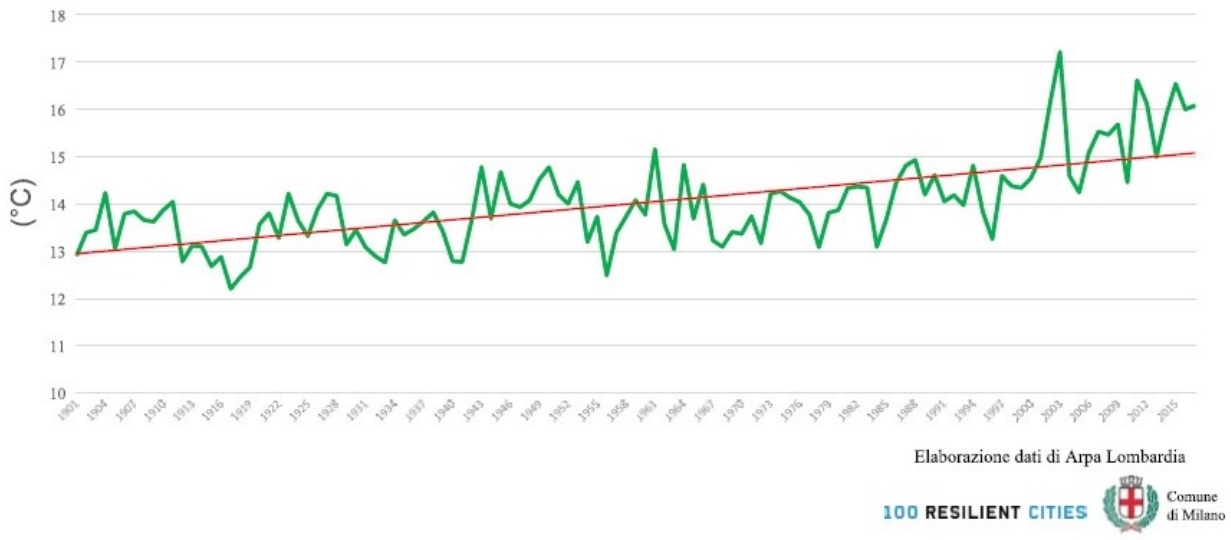


Figure 1: *Increase in average temperatures in Milan from 1901 to 2017,* ARPA *data: increase of 2°C*

It is therefore clear that a substantial presence of trees and green areas, which would provide shade to the city during the hottest hours and reduce the amount of asphalt, would make the city more livable in terms of sustainable and local mobility by bicycle or on foot, particularly for the most vulnerable groups or the elderly. On the other hand, there is also to consider that when thinking about 15-minute cities, one cannot only consider the placement of certain services in specific areas; attention must also be given to the actual ease of access to these services by the population. It is clear that a society that excludes or at least penalizes the most vulnerable social groups cannot be classified as smart. Reducing temperature and air pollution during the hottest months, and perhaps adding services such as public fountains with the creation of green areas, are fundamental aspects to truly realize a city of this type that is accessible to everyone.

Furthermore, in the perspective of countering the most negative and impactful implications of climate change on metropolises like Milan, trees and green areas play a crucial role in mitigating the effects of extreme weather events such as hailstorms, heavy rainstorms, and tropical phenomena, which have become more frequent in

recent years, and floods, whose number has reached 20 since 2010 between the Lambro and Seveso rivers [9]. Tree roots improve the soil's ability to absorb water, reducing the risk of flooding and allowing the city to better react naturally to downpours and heavy rains, thus counteracting or at least limiting their impact. The presence of trees can also reduce wind speed, limiting the damage caused by storms like the one in 2020 in Brescia and surroundings, characterized by winds reaching 130 km/h.

There is, however, a negative aspect that needs to be considered, not so much regarding the placement of trees in the city, but rather concerning green areas such as urban parks. In some more degraded and peripheral areas, indeed, what should be urban parks conceived as places of community aggregation, especially for children, become real drug dealing squares or gathering points for local petty crime, mainly consisting of marginalized individuals, non-EU immigrants, and those from difficult social backgrounds [10]. Examples include parks in areas like Rogoredo, San Donato, Porta Romana, or, more recently, Parco Candia [11]. The placement of parks in those areas, therefore, has the opposite effect of what is hoped for: families tend to avoid them, aware of the dangerous environment, and the surrounding area, instead of gaining economic value for the properties, will end up losing it, as these green areas increase the degradation and danger of the area, attracting petty criminals even from the surrounding areas.

1.3 Our objectives

Our project can be summarized within the field of urban informatics: the study of urban contexts based on data derived from the network of people within an urban context and from city infrastructures [12]. In our case, specifically, we will use data provided by infrastructures within the city. In particular, we will make use of air quality data recorded throughout 2023 by various air quality monitoring stations located in different and specific areas of Milan, focusing particularly on particulate matter concentrations, measured with both PM10 and PM2.5 indicators [13].
To do this, we use data from two different sources: from the first, freely available among Milan's open data [14], we obtain nine stations, while from the second [15], an additional eight, for a total of 17 air quality monitoring stations located in the municipality of Milan, even though this number will decrease during the preprocessing phase.

Based on the values obtained from the measurements taken by these stations in 2023, we will try to determine if there is a correlation between air quality and the amount of green areas and, on the other side, of buildings located around the monitoring station.
Thus, in addition to the data available from the stations, we will also need data on the green areas present throughout the entire perimeter of the city: it will indeed be essential to have a clear and general overview of the location and quantity of the green areas already present in the territory, to understand how strong the correlation with air quality is, both in terms of number and concentration of green areas at a certain distance from the station and at a certain geographical direction (north, south, west, east).
The same applies, though in the opposite direction, to buildings: by obtaining the location of buildings, we will be able to see if there is a clear inverse correlation between building density and air quality.
We will therefore understand how greenery and buildings affect air quality, also providing a detailed analysis of the impact they have when located at a certain distance from the measurement point.

Having a clear understanding of the current situation is fundamental to be able to optimize based on precise and data-driven considerations of the state of the art. To obtain these data, we use another source this time, OpenStreetMap [16], which allows us to have a clearer view of the city of Milan, and therefore to consider, exploiting mainly community and satellite data, the location of buildings, roads and natural areas.

Based on this data, extracting as much information as possible will allow us to obtain a detailed overview of the current city situation, focusing on our prediction and optimization objectives.

Our first goal will be to obtain, based on the locations of buildings and natural elements within the city's perimeter, a prediction of what the values of a new hypothetical air quality monitoring station might be if placed in a specific location within the city. In particular, it will be possible to determine whether it will record air quality values above or below the average. This will be useful for understanding which areas of Milan are the most livable from this perspective and, of course, linking forward to the second objective, where it is most urgent to intervene with the placement of new green areas.

The second objective, in fact, is similar, but it no longer concerns estimating values recorded by a new hypothetical station placed at a specific location in the territory. Instead, it involves the placement of a new hypothetical green area. By adding this green area, likely a park, we will observe the improvements in air quality based on new estimated hypothetical measurements 'adding' this additional tree basin to Milan.

However, throughout all this processes, special attention will always be given to the most degraded areas, needing to evaluate case by case whether the installation of new green areas can have more positive or negative effects, and acting accordingly in the choice of the type of greenery to be implemented in those areas.

2 Data Analytics, Preparation for the Models and Prediction

2.1 Data Exploration and Preparation

As mentioned before, we retrieve the data regarding the air quality stations from two distinct sources. In the end, we will have the following monitoring stations:

- via Pascal
- viale Marche
- via Senato
- Verziere
- *viale Liguria*
- *p.le Abbiategrasso*
- *Parco Lambro*
- *p.le Zavattari*
- *via Juvara*
- Centro
- Porta Vittoria
- Via Meda
- Lodi
- Piazza Leonardo
- Via Ripamonti Fine
- Forlanini
- Villapizzone

where the stations in the first column are those obtained from the first source, and those in the second column from the second source. While the stations in the first column, those from the first source, already had 'official' names, we assigned unofficial names to those from the second source to identify them. The last stations listed in the first column, those in italics, were preliminarily excluded from the analysis. The Via Juvara station, the last of these, was excluded because it has been inactive since 2007 and so it obviously has no readings for the entire year of 2023. The others, although active, do not have measurements for either PM2.5 or PM10. As they are not useful for our analysis, they will also not be considered further.
In the map below, we show the location of these stations (with the exception of the ones excluded) within the municipality of Milan. The stations from the first source are shown in blue, while those from the second source are shown in red. It is noteworthy that there are no overlaps, and the stations appear to be fairly distributed across the more central area of the municipality.

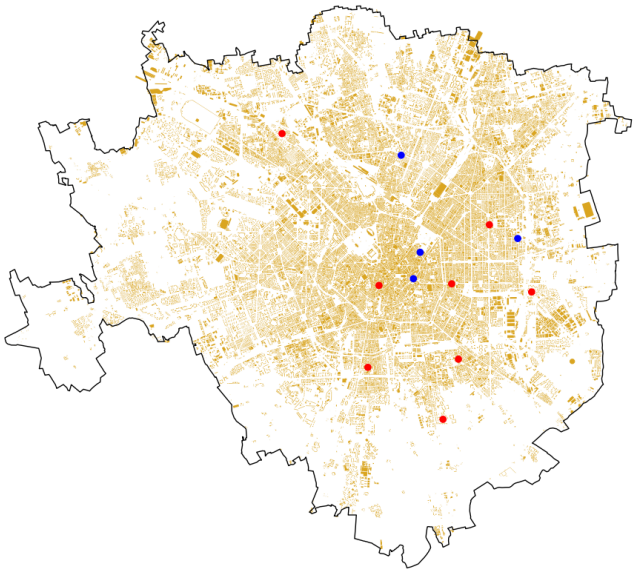


Figure 2: *Location of the stations within the boundaries of the municipality of Milan*

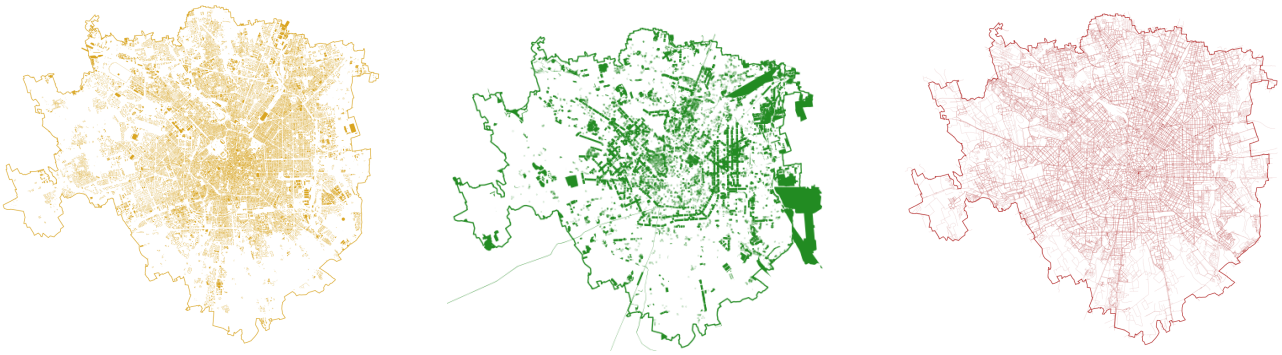
For these stations, coming from different sources, we naturally had different information. In the case of stations from the dataset provided by the Municipality of Milan, indeed, we had much more information compared to those from the second source: in addition to the measurements of particulates PM2.5 and PM10, which are common to both sources, we also had measurements of other chemical compounds, as benzene (C6H6), carbon

monoxide (CO), nitrogen dioxide (NO2), tropospheric ozone (O3), and sulfur dioxide (SO2).
During the data processing phase, for each station, we considered the measurements for each month of the calendar year 2023. Therefore, we reported the month, the pollutant substance, the number of measurements taken in that month for that substance, and the average value that it assumes within that month. Substances present only for some stations but not others were excluded, resulting in a final output like this:

Month	**PM10**	**PM10 measurements**	**PM2,5**	**PM2,5 measurements**
1	30.730758	13490	20.968428	13490
2	35.931124	13856	23.128715	13856
3	20.303564	12872	14.347259	12872
4	9.433302	6593	7.428934	6593
5	9.086172	17712	7.302746	17712

Table 1: Data of pollutant measurements for the first five month of 2023 recorded by the "Centro" station

It is worth to note that in the map before (figure 2), for better visualization, we have used the location of buildings within the municipality, shown in a dark yellow, as the background. This information, along with the location of natural elements, streets, and highways, was obtained through OpenStreetMap and thus from satellite data.
We therefore present below similar maps for the other elements:



(a) *Buildings* (b) *Natural elements* (c) *Streets*

Figure 3: *Locations of buildings, natural elements and streets in Milan, according to OpenStreetMap's data*

where in the case of the map that represents natural element we had to cut the map to refer only to the municipality of Milan, because elements like rivers (included in the water group) can indeed belong to Milan but also extend far beyond it.
However, for these reasons, we do not consider elements like rivers. Referring to the 'natural' column of OpenStreetMap data, after a preselection, we consider, as relevant for the continuation of our analysis, only natural elements categorized as trees, tree rows, grasslands, woods, scrubs, shrubs, tree groups and forests, for a total amount of 23 261 natural elements in Milan territory, number that we will increase as explained in Data Augmentation section.

As the final step in the data preparation phase, we also adjusted the CRS (Coordinate Reference System) of the geographic data. A CRS is a standardized system that defines a way to identify and position points on the Earth's surface. It can be divided into two types: geographic, which uses latitude and longitude ideal for representing positions on the Earth's spheroid but less suitable for distance calculations, and projected, which uses Cartesian coordinates and is essential for flat maps and distance calculations [17] [18].

Therefore, we converted our data from a geographic CRS (EPSG:4326) to a projected CRS commonly used for northern Italy (EPSG:32632), which corresponds to projection UTM zone 32N [19]. This will be useful for distance calculations, which are crucial for the next section.

2.2 Data Augmentation

Having obtained the list of active stations located in the area and their coordinates, we want to determine how much greenery is around each of them. Using the data on green areas obtained from OpenStreetMap, our goal is to obtain their coordinates and, using the coordinates of both the stations and the green areas, calculate the straight-line distances (i.e., Euclidean distances) between each station and each element of the green areas dataset.
Once the distances are calculated, for each station we consider concentric areas with the coordinates of that station as the common center, and increasing radii: $100\,m$, $200\,m$, $500\,m$, $1\,km$ and $2\,km$. This way, we obtain the amount of green area within these circles of increasing distance from the station, and we have no circle that extends beyond the perimeter of the city, for any station. But there is more: we consider that the station readings could also be influenced by winds and public buildings. It might indeed happen that the readings of a station benefit more from trees located to the north of it than to the south, if there are many tall buildings between the station and the trees to the south, which can impede proper air flow. Each of the circles created earlier will therefore be divided into 8 circular sectors of 45° each, with a unique numbering that starts from east and moves counterclockwise, following the notation used for angles when expressed in radians.

Moreover, when we record the data, the numbers will not be cumulative. This means that as we expand the radius of the sectors, we will never consider elements present in the smaller radius sector. So, we won't consider the sector with a 100-meter radius and the one with a 200-meter radius, but rather the sector with a 100-meter radius and the one with a radius from 100 to 200 meters. In this way, each point will be present in only one sector, reducing the risk of having concurrent features during the machine learning phase.

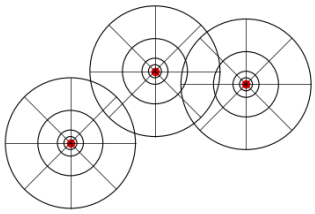


Figure 4: *Basic idea of the data augmentation model*

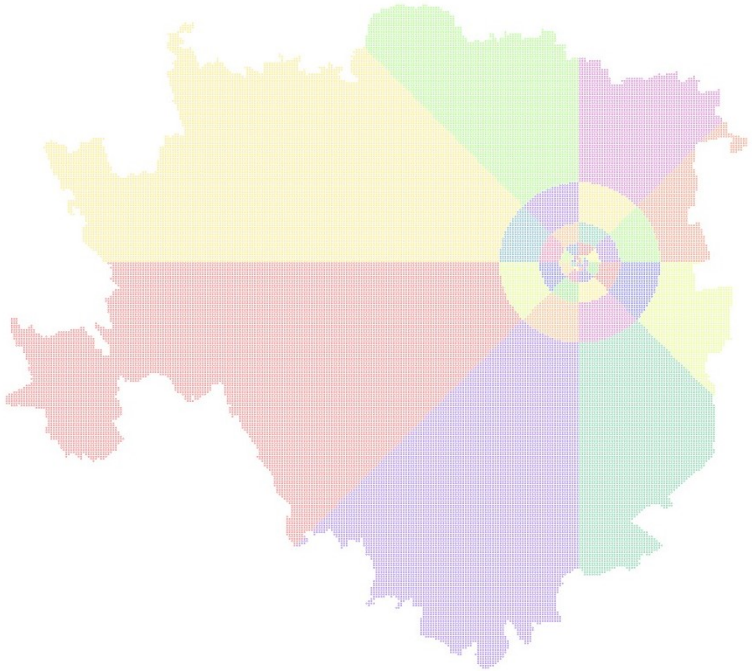


Figure 5: *The model applied and visible on a map of Milan, in this case with its center corresponding to the station in Piazza Leonardo*

However, there is a problem: as we saw during the data exploration phase, the green areas obtained from the OpenStreetMap dataset are divided into many categories, such as trees, woods, and 'grasslands'. While in the case of categories like trees, the coordinates are simply and intuitively expressed as a point in space, in the case of other categories, as woods or grasslands, the geometries used are the ones of Polygons, MultiPolygons, LineStrings and MultiLineStrings, according to the notation of GeoPandas.
Therefore, calculating the straight-line distance between trees and stations, that is, the Euclidean distance from point to point, is obviously straightforward; less so is the distance from a polygon to point, or from a line to a point either. Moreover, we would surely find ourselves in situations where a forest or a meadow would be straddling two sectors, if not three or four or more, at least according to the logic with which we defined the sectors earlier.
So, even reducing these areas to a single point at their center and calculating the distance would lose a lot of information if that area were distributed across multiple sectors. Additionally, reducing all areas to a single point regardless of their size would make us lose information about the size of the areas themselves: an area like Parco Nord cannot be reduced to a single point, as if it were just any tree or a small area.

Category	**Geometry/ies**	**Occurrencies**
tree	Point	19142
tree_row	LineString, MultiLineString, Polygon	2619
grassland	Polygon, MultiPolygon	1060
wood	Polygon, MultiPolygon	266
scrub	Polygon, MultiPolygon	147
shrub	Point	23
tree group	Polygon	3
forest	Polygon	1

Table 2: The categories we considered, sorted by descending order and with the corresponding geometry/ies reported

The approach we have chosen to follow is therefore as follows: for Polygons and MultiPolygons, we will develop a function that allows us to calculate their area. Once this is done, we will have a clear indication of their size. We will then create, with another function, a number of random points entirely within the polygon, and this number of points will be proportional to the area of the forest or meadow. We assumed that we consider one point for every 15 square meter, and in any case, at least one point per polygon (if the polygon has an area smaller than one, we still generate a point inside it). This value, 15, was chosen as it represents the second percentile of the areas of green elements classified as polygons. It is an arbitrary choice, which seemed to us a good compromise for achieving a sufficient density of points without distorting or falsifying the machine learning models, and the need not to create an excessively large number of points (already in this case we are dealing with hundreds of thousands, while for buildings, for which we will use a similar approach, we will exceed a million).
Therefore, we did not differentiate between various types of green areas, assuming that one square meter of grassland, for example, has the same "green density" as one square meter of forest. We expect the machine learning model to determine the greater importance of one type of area over another. With this approach, by the way, we will have reduced those areas to a set of points with specific coordinates, and for each point, we will calculate the Euclidean distance from the station, exactly with the same approach we use for the trees.
In the case of the other geometries, LineStrings and MultiLineStrings, the function does not calculate an area, but a length. Then, using a logic similar to that for areas, a number of points proportional to the length are generated in random positions along the segment. In this case, we assume to generate a point every two meters of length of LineStrings or MultiLineStrings.

The same identical approach has also been applied to buildings: for the same reasons, therefore, they were all converted into points, always calculating the area and generating a number, proportional to the value of that area, of random points within the area. For buildings, the area at the second percentile is approximately 32 square meters. Consistent with the approach for green areas, we will therefore generate one point for every 32 square meters.
In this case, however, we want the generated points to also be proportional to the building's height. Therefore, by considering both area and height, we make the number of points proportional to the volume of the building. The buildings dataset does indeed have a column indicating height, but it is filled with null values and is thus unusable. However, we have another very useful column: the number of floors in the building. Although this

column also contains some erroneous values or strings, and sometimes indicates floors with a float number (which were converted to their integer part), it is definitely usable. We calculate a multiplicative coefficient to consider in determining the number of points. This coefficient is defined as the integer part of the number of floors divided by three, plus one. Thus, if a building has 1 or 2 floors, the coefficient will be 1; if it has 3, 4, or 5 floors, it will be 2, and so on.

In the case of buildings, we deal only with Polygons and MultiPolygons, but in this case, there is a necessary data cleaning phase to apply. In GeoPandas, in fact, Polygons are simply defined as a sequence of coordinate points, where the points are the vertices of the polygon. Through simple Euclidean geometry, knowing the coordinates of the vertices, it is possible to calculate the sides and consequently the areas of the polygons. However, among our buildings, we found some - five - whose geometry consisted of polygons whose vertices all had the same identical coordinates, thus generating zero areas. In these cases, we simply treated them by automatically converting them into points using the corresponding coordinates, thereby skipping the area calculation.

However, computing the total area of all the green areas and of all the buildings before these processes, we obtain an area of approximately $17\,\mathrm{km}^2$ for the green spaces and of $32\,\mathrm{km}^2$ for the buildings. Therefore, in the city of Milan, buildings occupy almost twice the area compared to green spaces.

2.3 Machine Learning

At this point, our technique to feed into a machine learning model all the data we obtained in an appropriate manner is to aggregate all data by area and entity type. Summarizing what we did before, we now have 8 sectors (of 45 degrees each) and 5 distance rings, so we have a total of 40 areas; and, in addition to those, we have to consider them also for each of the 9 green area types, plus the building type. So, in total, we have $40 \times 10 = 400$ columns representing the surrounding areas with respect to each station. We also feed the normalized properties representing the position (x and y axis) to the model.
With respect to the air pollution data, we aggregate it by month (with a mean over the days), and then we encode the months as dummy variables. Finally, we add dummy variables for the dataset, since we cannot exclude a bias due to the hardware used by the different data sources.

Since our preliminary objective is to understand the link between our covariates and pollution, we exclude in principle dimensionality extraction techniques before the ML training phase. For the same reason of interpretability, we are also looking for a model that is not opaque, in the sense that we want them to provide clear results and highlight positive and negative correlations, as well as their intensity.
Thus, the models we used for our task are all Linear: we considered Least Squares Regression, Ridge, Lasso, and ElasticNet.
We used the first categorical dummy for each category group as the baseline category (resulting in 'January' and 'station obtained by Milano data'). For this task, we manually checked the coefficients of each covariate to draw an accurate picture of the weight of all the columns. The target variable we chose is PM10, but this experiment is replicable with other indicators without modifications.

To choose a good model for the prediction task, instead, we added to the evaluation list the tree-based ones, but most of all we used Lasso feature reduction, which drastically reduced the features to 16, from the more than 400 columns we had in the beginning. Before the feature reduction phase, we also made a tentative with a new level of augmentation: sum each entity area values per ring or per sector (thus considering only just the angle or just the distance, and not both together). We compared the models using 10-fold cross-validation measuring the R^2 value. The analysis of the R^2 values and the subsequent selection of the best model to be used for the next steps is deferred to the results section.

With the best model obtained, in fact, we moved forward to the more advanced objectives of the project. We proceeded to re-pre-process all the buildings and green areas data with respect to a new point to estimate the pollution there. We chose to consider a point in the Darsena as the location for our new hypothetical station. The Darsena is an area in Milan near Porta Ticinese, famous for its nightlife, with many people frequenting the area. Therefore, they may be interested in assessing the air quality and livability. Currently, there are no sensors in that area, or at least not with available data. Note also that we do not have data about the western part of the city: this prediction can also be useful to cover new areas, not just estimate pollution in points of interest.
Unfortunately, we could not proceed in estimating the Milan area pollution point by point (for example every 200 meters) due to computational limitations and also due to the fact that, for points near the city boundary, we would have a bias due to the fact that our model does not consider buildings and green areas outside the

city's territory, which would clearly have an influence in those cases; but, anyway, our theoretical framework is also ideated to be used in such objectives.

Now, the idea is to introduce a new green area within the city, currently non-existent and of significant size, and to observe the effects on the predicted pollution values recorded by the stations. Obviously, since the stations, in our model, are influenced by elements within a maximum distance of 2 km, the insertion of a new green area in a specific part of the city cannot affect all the stations. To decide where to place the area and its dimensions, and to avoid placing it in zones that might not benefit socially, we draw inspiration from an existing project: the creation of a 300 000 square meter park in the Scalo Farini area [20]. To create it, since we do not know the actual boundaries of the park according to the real project, we take the coordinates of Lampo Scalo Farini and consider them as the center of a circular area of 30 000 square meters that we create specifically. This area will then undergo the usual data augmentation process, with the creation of points within it, classified as grasslands. In this scenario, only one station from our data (Viale Marche) can be impacted by this new green area. Exactly like the pollution city overview at the end of the previous paragraph, with this technique it is possible to draw a picture of the pollution in Milan after the completion of all the 18 projects involving each one various green areas, but we don't have enough computational power.

3 Results

The analysis of the different linear regression coefficients highlighted the good importance of the periods of the years, expressed as months (for example, the October coefficient on average reduces PM10 levels by 3 whole points with respect to the baseline January), and a lower impact from the normalized coordinates and data source (Lasso and ElasticNet put their coefficients to 0, removing them). Regarding the areas, the results are mixed: there are some coefficients to 0 due to not enough data: for example a wood never appeared in the eastern area of a station between 100 and 200 meters, or wetlands never appeared at all, because they were all too far from any reference station). There are also some green areas increasing pollution and coefficients clearly too strong, but many have the values in the right direction. We found by querying the training data that the cause of this seems to be the too-high level of granularity we took into account, and, in this first phase before feature selection, of course, collinearity. For example, in least squares we have that trees in sector 1 within 100 meters from the station have a great beneficial effect on pollution, but sector 2 within 100 meters has an effect similar to the opposite; of course, the different models do not agree on this position, but there is no really stable model nonetheless.

After these findings we repeated the experiment ignoring the sectors, meaning that now the covariates are not an intersection of entity and angle and distance, but just entity and distance. We discarded the angle option with feature selection. The results for months, dataset and coordinates are in line with the previous, while the coefficients for the rings are now much better, although there are still some negative-coefficient buildings and positive-coefficient green areas. For every entity, the coefficients go from an impact of 10^{-5} for the nearest areas reaching 10^{-3}, usually for the 500-1000 m or the 1000-2000 m rings. This indicates that our way of dividing the entities by distance with a ring width that follows a geometric progression was good, but it would have been slightly better to have larger rings near the station and probably to consider rings more far than 2km (but we could not for computational reasons). A Great example is the 2000m ring building coefficient being 0.0018, meaning that each 15 m^2 increases PM10 by around that value (15 000 m^2 of buildings, around the size of two soccer fields, increases PM10 by almost 2 points). The most positive impact is given by grasslands and tree rows. Considering the 5 times difference of points number for trees in preprocessing, also the trees are important but a bit less. The results are as expected, and the details for all the selected covariates after feature selection for the least square regression trained again on the whole dataset can be found in table 3 .

Regarding the best model, with the Lasso feature selection and some manual adjusting the most important dimensions are identified as months, buildings, grasslands, woods and tree rows. All the choices are in line with our expectations, except for the trees we believed to be included in the list: it seems that single tree points, although more in number, have less impact compared to more extended or dense green areas. Among the linear models plus the decision tree, gradient boosting (trees), KNN regression, and linear SVM, the best is the Gradient Boosting with an R^2 of around 0.5. Details are shown in table 4.

Covariate	**Value**	**Unit**
Intercept	26.55258914471826	PM10
month_2	9.960776	PM10
month_3	-11.990735	PM10
month_4	-21.803305	PM10
month_5	-22.936624	PM10
month_6	-21.728732	PM10
month_7	-20.854082	PM10
month_8	-23.906450	PM10
month_9	-18.140619	PM10
month_10	-8.614080	PM10
month_11	-12.729556	PM10
month_12	1.400593	PM10
buildings_100	-0.000099	PM10 / 32m^2 of buildings
buildings_200	-0.000503	PM10 / 32m^2 of buildings
buildings_500	0.001901	PM10 / 32m^2 of buildings
buildings_1000	-0.000595	PM10 / 32m^2 of buildings
buildings_2000	0.001877	PM10 / 32m^2 of buildings
green_grassland-100	-0.000443	PM10 / 15m^2 of grassland
green_grassland-200	-0.001513	PM10 / 15m^2 of grassland
green_grassland-500	-0.002991	PM10 / 15m^2 of grassland

green_grassland-1000	-0.000783	PM10 / $15m^2$ of grassland
green_grassland-2000	0.000074	PM10 / $15m^2$ of grassland
green_wood-1000	-0.000340	PM10 / $15m^2$ of wood
green_wood-2000	-0.000080	PM10 / $15m^2$ of wood
green_tree_row-1000	-0.002526	PM10 / $15m$ of consecutive trees
green_tree_row-2000	-0.001053	PM10 / $15m$ of consecutive trees

Table 3: The regression coefficients of Linear Least Square regression, where PM10 unit of measure stands for $\mu g/m^3$ of PM10

Model	R^2
Gradient Boosting	0.487
Linear Least Squares	0.235
Linear Ridge	0.161
Decision Tree	-0.103
Linear Lasso	-0.304
SVM (linear)	-0.383
Linear ElasticNet	-0.623
KNN	-13.9

Table 4: The value of R^2 for each model

Lastly, we successfully found that the Scalo Farini would lower the PM10 values in Viale Marche by 0.16 points, from a value of around 30.18 $\mu g/m^3$ to 30.02 $\mu g/m^3$, which is absolutely significant considering the 2 kilometres distance. And we found that in Darsena the actual level of pollution for each month are, in order of month:

Month	Value, in $\mu g/m^3$
January	25.255
February	46.958
March	23.105
April	13.287
May	12.737
June	13.323
July	13.147
August	12.528
September	16.059
October	26.450
November	22.194
December	40.471

Table 5: The level of pollution for each month

So, their average is 22.13 $\mu g/m^3$, a little below the Milan average of our available stations of 22.37 $\mu g/m^3$.

4 Future Developments

As future developments, an idea would certainly be to also leverage road data, which, as shown in the data exploration phase, was also easily obtainable from OpenStreetMap but was not utilized. A similar approach could be applied as with green areas and buildings for data augmentation, generating points to represent the road network. However, it would be crucial to integrate data on the location of the road network with traffic data to understand which roads have the greatest impact on pollution and act accordingly. Additionally, as the R^2 shows, our model does not take into account a not so small part of the variability of the data: such data integrations, adding for example also the weather, will increase the score for sure.

Another absolutely interesting analysis would be a parallel study to that already conducted on Scalo Farini Park, but with a different approach: it would no longer be about observing the effects resulting from an existing project. Instead, given an available green area of, for example, 1000 square meters, the goal would be to determine which area of Milan would benefit most from inserting this hypothetical green area. In doing so, we must consider the issue of degraded areas and think about the effects, not only environmental but also socio-economic, that the placement of a park in certain zones could entail.

References

[1] GeoPandas documentation
https://geopandas.org/en/stable/docs.html

[2] Laura Sartori and Davide Arcidiacono: *In Search for (the Lost) Smartness in the Evolution of the Smart Cities: Consumers or Citizens?*
http://hdl.handle.net/10125/70917

[3] United Nations Climate Change: *What is the Kyoto Protocol?*
https://unfccc.int/kyoto_protocol

[4] European commission: *Covenant of Mayors - Objectives and Key Pillars*
https://eu-mayors.ec.europa.eu/en/about/objectives-and-key-pillars

[5] Simon Elias Bibri and John Krogstie: *Smart sustainable cities of the future: An extensive interdisciplinary literature review*:
https://www.sciencedirect.com/science/article/pii/S2210670716304073

[6] Fulvia Pinto and Mina Akhavan: *Scenarios for a Post-Pandemic City: urban planning strategies and challenges of making "Milan 15-minutes city*:
https://www.sciencedirect.com/science/article/pii/S2352146521009480

[7] Comune di Milano: *Milan 2020 Adaptation strategy -Open document to the city's contribution*:
https://www.comune.milano.it/documents/20126/7117896/Milano+2020.+Adaptation+strategy.pdf/d11a0983-6ce5-5385-d173-efcc28b45413?t=1589366192908

[8] Corrado Fontana - Valori: *Milano, +4 gradi nel 2050. Contro la crisi climatica, ecco "l'urbanismo tattico"*
https://valori.it/milano-crisi-climatica-urbanismo-tattico/

[9] Giovanna Maria Fagnani - Corriere della Sera: *Così cambia il clima di Milano: 2,1 gradi in più delle temperature in dieci anni, tempeste ed esondazioni*
https://milano.corriere.it/notizie/cronaca/21_novembre_23/cosi-cambia-clima-milano-21-gradi-piu-temperature-dieci-anni-tempeste-ed-esondazioni-ed7e2136-4c61-11ec-93ad-d9e7f28c53fe.shtml?refresh_ce

[10] Elisabetta Santon - TGR Lombardia: *Tra Rogoredo e San Donato, nella terra di nessuno dello spaccio di droga*
https://www.rainews.it/tgr/lombardia/video/2023/09/spaccio-droga-san-donato-rogoredo-milano-33534a3f-f4dc-434b-a4bf-280c47168103.html

[11] Massimiliano Mingoia - Il Giorno: *Tossici e sbandati al parco Candia, davanti ai bambini. La protesta delle famiglie: "Qui è la stanza del buco"*
https://www.msn.com/it-it/notizie/milano/tossici-e-sbandati-al-parco-candia-davanti-ai-bambini-la-protesta-delle-famiglie-qui-%C3%A8-la-stanza-del-buco/ar-BB1nQkEF?ocid=BingNewsSerp

[12] Foth, Marcus and Choi, Jaz Hee-jeong and Satchell, Christine: *Urban informatics*
https://doi.org/10.1145/1958824.1958826

[13] ARPA Lombardia: *PM10 E PM2.5*
https://www.arpalombardia.it/temi-ambientali/aria/inquinanti/pm10-e-pm25/

[14] Comune di Milano - Stazioni monitoraggio inquinanti - *last update: 30/04/24*
https://dati.comune.milano.it/dataset/ds484_stazioni_di_monitoraggio_inquinanti_atmosferici_dellarpa_sit

[15] Sensor Community: - *last update: 20/06/24*
https://sensor.community/en/

[16] OpenStreetMap - Comune di Milano
https://www.openstreetmap.org/relation/44915

[17] Sistema di Riferimento delle Coordinate - QGIS documentation
https://docs.qgis.org/3.34/it/docs/gentle_gis_introduction/coordinate_reference_systems.html

[18] Geopandas documentation - Projections
https://geopandas.org/en/stable/docs/user_guide/projections.html

[19] MapTiler - epsg.io
https://epsg.io/32632

[20] Milano Città Stato - I 18 grandi progetti di rigenerazione che trasformeranno Milano e il suo Hiterland
https://www.milanocittastato.it/milano/lacittadeisogni/i-18-grandi-progetti-di-rigenerazione-che-trasformeranno-milano-e-il-suo-hinterland/