Jan, 2025

GRINS DISCUSSION PAPER SERIES DP N° 126/2026 ISSN 3035-5576

GRINS

Statistical methods for evaluating the environmental sustainability of urban infrastructure

DP N° 126/2026

Authors:
Andrea Gilardi, Francesca Ieva, Laura Sangalli, Piercesare Secchi

Jan, 2025
GRINS DISCUSSION PAPER SERIES DP N° 126/2026 ISSN 3035-5576

Statistical methods for evaluating the environmental sustainability of urban infrastructure

Andrea Gilardi, Francesca Ieva, Laura Sangalli, Piercesare Secchi

KEYWORDS

Environmental sustainability | PM10 emissions | Urban infrastructure | Air pollution risk mapping

JEL CODE

Q53, C38, R11

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Title: Statistical methods for evaluating the environmental sustainability of urban infrastructure. Publication Date: 2025.

This study employs an original statistical method to assess the environmental sustainability of urban infrastructure in terms of PM10 emissions. The methodology uses the data produced by the monitoring stations and, through principal component analysis and interpolations, yields as an outcome the probability that a certain municipality exceeds the PM10 threshold, providing a detailed map of areas with the highest risk of air pollution.

Statistical methods for evaluating the environmental sustainability of urban infrastructure

Andrea Gilardi[1], Francesca Ieva[1], Laura Sangalli[1], Piercesare Secchi[1]

[1]*Polytechnic University of Milan, Department of Mathematics*

January 7, 2025

Abstract

This study employs an original statistical method to assess the environmental sustainability of urban infrastructure in terms of PM_{10} emissions. The methodology uses the data produced by the monitoring stations and, through principal component analysis and interpolations, yields as an outcome the probability that a certain municipality exceeds the PM_{10} threshold, providing a detailed map of areas with the highest risk of air pollution.

Keywords: Environmental sustainability; PM10 emissions; Urban infrastructure; Air pollution risk mapping

JEL codes: Q53; C38; R11

1. Introduction

PM_{10}, particulate matter with a diameter of 10 micrometers or less, is an air pollutant that poses significant environmental and public health challenges, particularly in urbanized and industrial regions. These fine particles, primarily generated by traffic, industrial activities, and residential heating, can penetrate deep into the respiratory system, contributing to a range of health problems, including respiratory infections, cardiovascular diseases, and premature mortality.

Several studies have consistently shown a strong association between PM_{10} exposure and adverse health outcomes, particularly in areas with high population density and industrial activities (Pope and Dockery 2006; Brunekreef and Holgate 2002). In Lombardy, one of the most industrialized and densely populated areas of Italy, PM_{10} levels frequently exceed both national and European air quality standards (Marcazzan et al. 2003; Cattaneo et al. 2011), despite a long-standing commitment to reducing air pollution levels. This has raised serious concerns about the long-term health implications for the region's inhabitants, making the analysis of PM_{10} data crucial for developing effective public health policies.
Analyzing PM_{10} data is not only important for tracking compliance with regulatory standards, but also for identifying seasonal and geographical patterns that influence air quality. Numerous studies have demonstrated that PM_{10} concentrations vary significantly across regions depending on external factors such as topography, orography, weather conditions, and human activities (Querol et al. 2004). In Lombardy, where industrial activity and heavy vehicular traffic are concentrated in urban centres, understanding these patterns is essential for formulating mitigation strategies.
In this document, we present a statistical analysis of PM_{10} quality data exploring a method to predict the probability that concentrations exceed a health-related threshold. The analysis starts from the estimation of PM_{10} density distributions at the monitoring station level, followed by a principal component analysis which was developed in the space of density functions to properly account for the compositional nature of the data. The principal component scores were then interpolated at the municipal level using a spatial interpolation technique known as block-kriging and showcasing a possible method for down-scaling air pollution data. The final step of the analysis consisted in computing the probability of exceeding a PM_{10} threshold at the municipal level, providing a detailed map of areas with the highest risk of air pollution.
The remainder of this document is structured as follows. We first provide an overview of the literature regarding air quality data analysis. Then, we describe the database used in our analysis and the statistical framework. We finally present the results and discuss possible policy implications.

2. Outline of literature background

Analysis of air quality data is a topic of great interest in environmental science and public health research. Several studies investigated the spatial and temporal variability of air pollution levels and their impact on human health and the environment. For example, Pope and Dockery (2006) conducted a comprehensive review of the health effects of air pollution, highlighting the association between particulate matter exposure and respiratory and cardiovascular diseases.

More recently, Chen and Hoek (2020) replicated their efforts and performed a systematic review of evidence of associations between long-term $PM_{2.5}$ and PM_{10} exposure in relation to all-cause and cause-specific mortality. The authors found that long-term exposure to particulate matter is associated with increased mortality from cardiovascular diseases, respiratory diseases, and lung cancer.

In the context of air quality data analysis, several statistical methods have been proposed during the past years to model and understand the spatial and temporal variability of air pollution levels. In particular, focusing on the spatial domain, Berrocal, Gelfand, and Holland (2010) applied spatio-temporal kriging (N. Cressie and Wikle 2011) to model PM_{10} concentrations combining observed data with output from numerical models. Their proposal represents a valuable tool to spatio-temporally downscale and predict air-pollution at a finer spatial scale. Fotheringham, Yue, and Li (2019) used Geographically Weighted Regression to examine the influences of air quality in China's cities, highlighting the importance of considering local factors when analysing air pollution data. We refer to Gelfand et al. (2019) for a comprehensive overview of statistical methods in environmental science.

More recently, also Functional Data Analysis (FDA) (Ramsay and Silverman 2005) tools have been applied to the analysis of air quality data. Ignaccolo, Mateu, and Giraldo (2014), for example, proposed a (universal) kriging approach to model spatially-distributed functional data (e.g. PM_{10} concentrations) taking into account exogenous variables related to meteorology, topology, and orography. Aguilera-Morillo, Durbán, and Aguilera (2017) proposed the adoption of penalized approach methods (as an alternative to kriging and kernel smoothing techniques) to predict the spatial distribution of functional variables.

Probability density functions represent a particular example of functional data (Egozcue, Dı´az–Barrero, and Pawlowsky–Glahn 2006) and an idea tool to evaluate the probability that air pollutants' concentrations exceed their regulatory constraints. Their statistical modelling is particularly challenging due to the (infinite) compositional nature of the data and their intrinsic constraint: $\int f = 1$ and $f \geq 0$. The literature is still at the beginning, but some recent contributions have proposed innovative methods to address these challenges. For example, Hron et al. (2016) proposed a method called Simplicial Functional Principal Component Analysis (SFPCA) that allows for the analysis of density functions in the Bayes Space, providing a powerful tool for extracting key patterns of variability from functional data. Their application in the context of air quality data analysis is still limited, but the potential of these methods is promising.

3. Data and descriptive statistics

The data provider considered for the analyses described in this study is the European Environment Agency (EEA), which is a supra-national institution that collects and harmonizes air quality measurements for a panel of 38 countries.[1] More precisely, the EEA shares[2] a database that includes both raw and validated air quality measurements obtained from geo-referenced time series data. The data exchange process between each member country and the EEA is handled by a partnership network named European Environment Information and Observation Network (EIONET).

After downloading the complete database from the aforementioned sources, we selected the PM_{10} measurements pertaining to the monitoring station located in the Lombardy region (Italy) during the time period 2018 - 2022. We decided to focus on PM_{10} since it is one of the most commonly measured air pollutants and it has well-known adversarial effects on human health and the environment. Therefore, it is a good candidate to showcase the potential of developing statistical analysis to study environmental and health-related issues. The Lombardy region was selected as a first case study since it is characterized by high population density (especially in the areas near Milan and other parts of the Po Valley), high level of industrialization, and a mix of urban, rural, and mountainous territories. Finally, we focused our efforts on the period 2018 - 2022 since it is, in our opinion, a sufficiently long time period to collect a meaningful amount of data for a statistical approach, but not too long to confound the analyses, mixing long-term trends with short-term fluctuations.

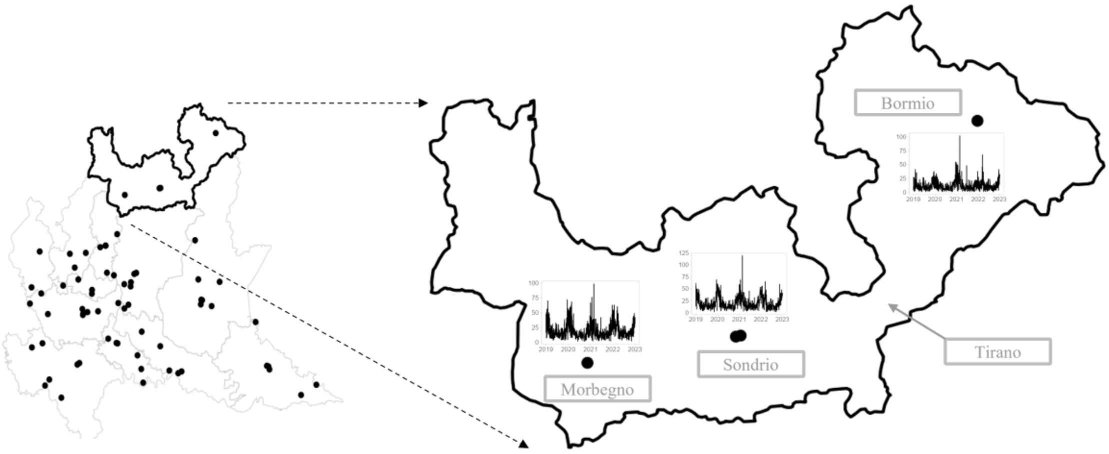
The spatial distribution of the monitoring stations recording PM_{10} in Lombardy during the period under analysis is displayed in Figure 7.1 (left) as black dots. As we can see, the stations are spread in several parts of the region, with a higher concentration near the most important cities like Milan, Bergamo, or Brescia. The right part of Figure 7.1 zooms on the province of Sondrio. Each dot represents a monitoring station, whereas the (spatially-referenced) time series represent the PM_{10} concentration recorded by the closest station. The time series shows that PM_{10} is higher in the winter months and lower in the summer months, which is a well-known pattern due to the increased use of heating systems during colder times. Similar behavior is observed in the other provinces of the region.

---

[1] The 38 countries are collectively named EEA38 and the complete list can be browsed at the following link: https://land.copernicus.eu/en/faq/general-questions/what-is-eea38

[2] The air-quality database created and shared by the EEA, together with the corresponding meta-data, can be browsed at the following link: https://www.eea.europa.eu/en/datahub/datahubitem-view/3b390c9c-f321-490a-b25a-ae93b2ed80c1

Figure 7.1. Monitoring stations recording PM_{10} in Lombardy in 2018-2022 (left) and zoom on the province of Sondrio (right).

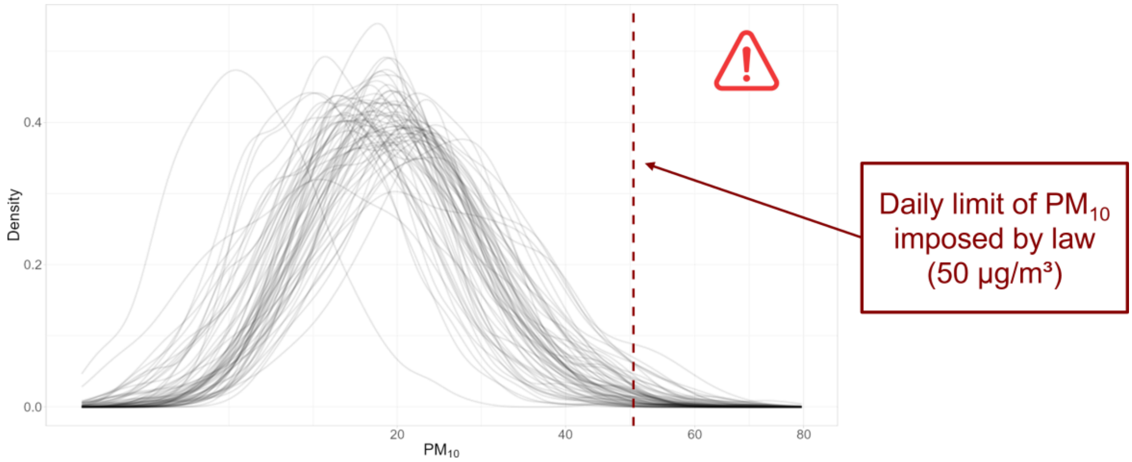


As already mentioned, the objective of our study was to develop a statistical analysis that evaluates the environmental sustainability at the municipal level. As a case study, we decided to focus on the PM_{10} pollutant and predict the probability that, at a certain point, the PM_{10} concentration exceeds a health-related threshold set by local and national authorities. In particular, we focused on the PM_{10} threshold set by the Italian legislative decree n. 155/2010[3], which is equal to 50 $\mu g/m^3$ and must not be exceeded more than 35 times in a year.

The practical computation of such probabilities of exceedance requires a shift in the data representation, going from (spatially-referenced) time series, such as those displayed in Figure 7.1, to (spatially-referenced) density functions. In fact, a density function is a functional data object that represents the probability distribution of a set of data (e.g. air pollution measurements). We computed the PM_{10} density distribution for each monitoring station in Lombardy via Kernel Smoothing methods (Silverman 1986), considering only the measurements recorded during Spring and Summer months. The output is displayed in Figure 7.2. As we can see, the density curves are reasonably symmetric around the mean value (which is around 20 $\mu g/m^3$), but they are sometimes characterized by long right tails, especially in stations located near industrial areas in the Po Valley. The vertical red dashed line denotes the threshold value of 50 $\mu g/m^3$ set by law.

[3] The complete text regarding the legislative decree can be browed at the following link: https://www.gazzettaufficiale.it/eli/id/2010/09/15/010G0177/sg. In particular, we refer to Allegato XI.

Figure 7.2. Density of the PM_{10} concentration for each monitoring station in Lombardy in 2018-2022, considering only spring and summer. The density curves are computed via Kernel Smoothing methods



4. Methodology

A proper analysis of PM_{10} density curves must take into account the particular nature of the data. In fact, density curves are not only functional observations, but they are also strictly positive functions that integrate to 1 over their domain. Traditional methods (such as functional logistic regression) do not account for these characteristics and may lead to incomplete or biased conclusions. Therefore, we adopted an ad-hoc approach whose theory is rooted in the so-called Bayes Space, hereby denoted as B^2 (Egozcue, Dı´az–Barrero, and Pawlowsky–Glahn 2006).

The Bayes Space represents a generalization of the Aitchison geometry (Aitchison 1982) to density functions, which is crucial for ensuring that statistical analyses respect the compositional nature of the data and their constraints (Egozcue, Dı´az-Barrero, and Pawlowsky–Glahn 2006). Furthermore, the Bayes Space, equipped with a certain definition of sum, inner product, and product-by-constant, is a Hilbert Space and, as such, it is isomorphic to L^2 (the space of square-integrable functions) (Van den Boogaart, Egozcue, and Pawlowsky-Glahn 2014). This isomorphism allows us to perform standard operations (e.g. functional regression or functional principal component analysis) on B^2 in a statistically sound manner via the following pipeline:

1. Transform the data from B^2 to L^2 via the centred log-ratio (clr) transformation.
2. Perform the statistical operation in L^2 using, for example, the techniques described in Ramsay and Silverman (2005).
3. Transform the results back to B^2 via the inverse-clr transformation.

We employed the aforementioned strategy to develop a statistical model that predicts the probability of exceeding the PM_{10} threshold using a technique known as Simplicial Functional Principal Component Analysis (SFPCA) (Hron et al. 2016). The SFPCA allows us to extract the functional principal components (FPC) and the corresponding scores of the density functions in the L^2 space. By construction, the (SFPCA) scores are real numbers representing the projection of the original density curves onto the FPC. Considering the spatial dependence among the PM_{10} values recorded by proximate monitoring stations and the need to predict probabilities at the municipal level, we projected the SFPCA scores on the municipality grid using a spatial interpolation technique known as block kriging (N. A. C. Cressie 1993). Finally, we rebuilt the density curves from the kriged scores and computed the probability of exceeding the PM_{10} threshold.

The complete pipeline of the analyses is detailed as follows:

4. Compute the density curves for each monitoring station in Lombardy via Kernel Smoothing methods.
5. Transform the density curves from B^2 to L^2 via the clr transformation.
6. Perform the SFPCA on the clr-transformed density curves. The analyses summarized in the remaining part of this document were performed considering 3 FPC.
7. Krige the SFPCA scores at the municipal level to obtain a map of values all over the Lombardy region.
8. Build the clr-transformed curves from the kriged scores.

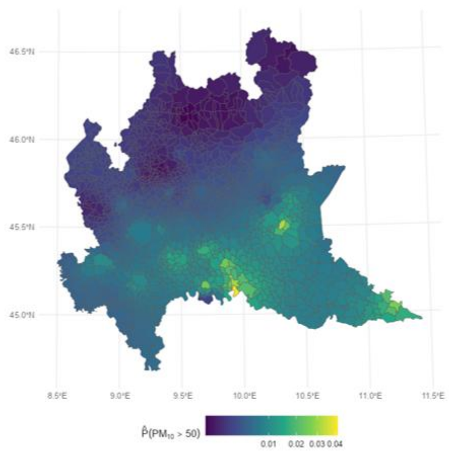
Transform the clr-transformed curves back to B^2 via the inverse-clr transformation and compute the probability of exceeding the PM_{10} threshold of 50 µg/m3.

5. Findings and discussion

The application of SFPCA to the PM_{10} density functions revealed strong spatial dependencies among the FPC scores. These scores provide a compressed representation of the data, highlighting key patterns of variability across Lombardy. Notably, the spatial dependence indicates that certain areas of Lombardy are consistently subject to higher PM_{10} levels, while others experience more modest fluctuations. The analysis also demonstrates that the scores corresponding to the three principal components are almost uncorrelated, indicating that the dominant sources of variability in PM_{10} levels are distinct from one another. This suggests that different geographical or environmental factors may be driving these patterns, such as proximity to industrial activities, traffic density, or topographical features.

As already mentioned, we applied block kriging to reconstruct the PM_{10} densities across Lombardy for areas without direct monitoring. The most relevant outcome of the analysis is the estimation of the probability of exceeding the PM_{10} threshold and the results are displayed in Figure 7.3. As we can see, the highest probabilities are located in the Po Valley, near the most industrialized and flat areas. In contrast, the northern parts of the region (which are characterized by mountainous areas) exhibit lower probabilities, reflecting the lower population density and industrial activity in these areas, but also the combined effect of the topology and orography which influence the dispersion of pollutants.

Figure 7.3. Predicted probability of observing a PM_{10} concentration higher than 50μg/m^3 at the municipal level in Lombardy. The probabilities are computed via the application of Simplicial Functional Principal Component Analysis and block kriging methods



6. Policy implications and recommendations

The results of our analysis may have important implications for environmental policy and public health in Lombardy. By identifying areas and municipalities with the highest probability of exceeding the PM_{10} threshold, we can help local authorities target interventions to reduce air pollution and protect public health. For example, local governments could implement traffic management measures, such as low-emission zones or congestion charges and, similarly, urban planning policies could be adjusted to reduce the impact of industrial activities on air quality in these areas.

To conclude, we point out that the spatial interpolation method considered in this document represent an example of down-scaling method which might be particularly valuable for predicting air pollution levels in smaller municipalities or rural areas where data coverage is limited. By combining these instruments with additional statistical analyses that estimate the uncertainty of the PM_{10} predictions, we might also provide local authorities with a toolkit to 1) assess the effectiveness of current policies, 2) evaluate the potential impact of new interventions, and 3) optimize the placement of new monitoring stations (see, for example, Zimmerman and Buckland (2019) and Zidek and Zimmerman (2019) and references therein).

7. Key references

[1] Aguilera-Morillo, M Carmen, Marı ́a Durbán, and Ana M Aguilera. 2017. "Prediction of Functional Data with Spatial Dependence: A Penalized Approach." Stochastic Environmental Research and Risk Assessment 31: 7–22.

[2] Aitchison, John. 1982. "The Statistical Analysis of Compositional Data." Journal of the Royal Statistical Society: Series B (Methodological) 44 (2): 139–60.

[3] Berrocal, Veronica J, Alan E Gelfand, and David M Holland. 2010. "A Spatio-Temporal Downscaler for Output from Numerical Models." Journal of Agricultural, Biological, and Environmental Statistics 15: 176–97.

[4] Brunekreef, Bert, and Stephen T Holgate. 2002. "Air Pollution and Health." The Lancet 360 (9341): 1233–42.

[5] Cattaneo, Andrea, Carlo Peruzzo, Gaetano Garramone, Patrizia Urso, Roberto Ruggeri, Paolo Carrer, and Domenico Maria Cavallo. 2011. "Airborne Particulate Matter and Gaseous Air Pollutants in Residential Structures in Lodi Province, Italy." Indoor Air 21 (6): 489–500.

[6] Chen, Jie, and Gerard Hoek. 2020. "Long-Term Exposure to PM and All-Cause and Cause-Specific Mortality: A Systematic Review and Meta-Analysis." Environment International 143: 105974.

[7] Cressie, Noel A. C. 1993. Statistics for Spatial Data. Revised. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119115151.

[8] Cressie, Noel, and Christopher K Wikle. 2011. Statistics for Spatio-Temporal Data. John Wiley & Sons.

[9] Egozcue, Juan José, José Luis Dı ́az–Barrero, and Vera Pawlowsky–Glahn. 2006. "Hilbert Space of Probability Density Functions Based on Aitchison Geometry." Acta Mathematica Sinica 22 (4): 1175–82.

[10] Fotheringham, A Stewart, Han Yue, and Ziqi Li. 2019. "Examining the Influences of Air Quality in China's Cities Using Multi-Scale Geographically Weighted Regression." Transactions in GIS 23 (6): 1444–64.

[11] Gelfand, Alan E, Montserrat Fuentes, Jennifer A Hoeting, and Richard Lyttleton Smith. 2019. Handbook of Environmental and Ecological Statistics. CRC Press.

[12] Hron, Karel, Alessandra Menafoglio, Matthias Templ, Klára Hrzová, and Peter Filzmoser. 2016. "Simplicial Principal Component Analysis for Density Functions in Bayes Spaces." Computational Statistics & Data Analysis 94: 330–50.

[13] Ignaccolo, Rosaria, Jorge Mateu, and Ramon Giraldo. 2014. "Kriging with External Drift for Functional Data for Air Quality Monitoring." Stochastic Environmental Research and Risk Assessment 28: 1171–86.

[14] Marcazzan, GM, M Ceriani, G Valli, and R Vecchi. 2003. "Source Apportionment of PM10 and PM2. 5 in Milan (Italy) Using Receptor Modelling." Science of the Total Environment 317 (1-3): 137–47.

[15] Pope, C. Arden, and Douglas W. Dockery. 2006. "Health Effects of Fine Particulate Air Pollution: Lines That Connect." Journal of the Air & Waste Management Association 56 (6): 709–42.

[16] Querol, Xavier, A Alastuey, CR Ruiz, B Artiñano, HC Hansson, RM Harrison, E ten Buringh, et al. 2004. "Speciation and Origin of PM10 and PM2. 5 in Selected European Cities." Atmospheric Environment 38 (38): 6547–55.

[17] Ramsay, J. O., and B. W. Silverman. 2005. Functional Data Analysis. 2nd ed. Springer New York, NY. https://doi.org/10.1007/b98888.

[18] Silverman, Bernard W. 1986. Density Estimation for Statistics and Data Analysis. 1st ed. Chapman; Hall.

[19] Van den Boogaart, Karl Gerald, Juan José Egozcue, and Vera Pawlowsky-Glahn. 2014. "Bayes Hilbert Spaces." Australian & New Zealand Journal of Statistics 56 (2): 171–94.

[20] Zidek, James V, and Dale L Zimmerman. 2019. "Monitoring Network Design." In Handbook of Environmental and Ecological Statistics, 499–522. Chapman; Hall/CRC.

[21] Zimmerman, Dale L, and Stephen T Buckland. 2019. "Environmental Sampling Design." In Handbook of Environmental and Ecological Statistics, 181–210. Chapman; Hall/CRC.

Final considerations and conclusions

Accessibility – the capacity of a city or territory to ensure that users can easily and effectively reach their destinations – is fundamental to the well-functioning of the socio-economic life of cities and territories. Accessibility makes a territory attractive for economic actors and tourists and contributes to the well-being of people. Accessibility is also crucial to reduce traffic, congestion and emissions, thus contributing to the environmental sustainability. To make cities resilient, accessibility has to be guaranteed also when the territory is hit by disruptions and shocks.

This report focuses on accessibility in general, and on its relationship with mobility in particular. The contributions presented in this report improve our knowledge of this relationship, by developing indicators and methodologies of analysis, and discussing empirical results and case studies. This knowledge can support urban and transport planners in developing solutions, investments and policies.

In particular, the policy recommendations that can be drawn from the contributions contained in this report can be summarized as follows.

First, any policy to improve accessibility should be targeted. As this report shows, there are tremendous differences across Italian territories in terms of accessibility and emissions, and also within the same city or area there are roads or infrastructure that are critical to ensure resilience.

Second, mobility policies should be inspired by a principle of integration of different modes. In particular, the integration between the public transportation system and the new forms of shared mobility seems to be an effective way to guarantee accessibility even when negative shocks occur. A promising solution to guarantee a successful integration is the realization of mobility hubs.

Third, policies should take into account that accessibility has significant welfare and equity implications. As discussed in the study on the health system in Rome, a small and otherwise negligible reduction in accessibility may be fatal when it pertains to a hospital.

Fourth, any policy to sustain accessibility should be based on a deep knowledge of the peculiar characteristics of the territory, of the existing network and of the potential users. In this respect, this report identifies indicators, statistical methods and strategies to exploit the huge amount of available data (or that can be collected through ad-hoc surveys). Again, integration of diverse data sources is crucial to achieve a more accurate and comprehensive picture of the phenomenon at hand. Moreover, new technology allow to collect information on the status of the mobility system in real time, and this information should be used by operators to continuously monitor accessibility and quickly identify and address issues, such as delays or service disruptions.