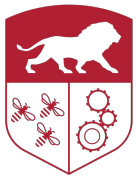
Politecnico di Torino



CENTRALE LYON

Politecnico di Torino

& École centrale de Lyon

Master's Degree in Mathematical Engineering

July 2025

Development of a Data-Driven Surrogate Model for Street-Scale Prediction of NOx Passive Scalar Concentrations in the San Salvario District of Turin

Supervisors:

Ridolfi Luca

Fellini Sofia

NGuyen Chi Vuong (École Centrale)

Salizzoni Pietro (École Centrale)

Candidate:

Kiefer Riccardo

Abstract

Clean air is essential to our health and to the environment. However, due to human activities causing polluting emissions, air quality has deteriorated considerably in the last century. These activities are notably linked to industry, energy production, domestic heating, agriculture and transport.

Air pollution is the number one environmental health problem in the EU. It causes serious illnesses such as asthma, cardiovascular problems and lung cancer, and vulnerable groups are affected the most. Air pollution also damages the environment and ecosystems through excess nitrogen pollution and acid rain. It is also costly for our economy, as it leads to lost working days and high healthcare costs.
Data from the European Environmental Agency reports 240,000 premature deaths caused by fine particulate matter annually and an annual economic cost of air pollution ranging from €231 to €853 billion.

Since the 1980s, the EU has been implementing policies on air quality that have contributed to a substantial decrease in most air pollutants over the past decades. To tackle air pollution and achieve the EU's zero pollution vision for 2050, the EU has a comprehensive clean air policy based on three pillars: ambient air quality standards, reducing air pollution emissions, and emissions standards for key sources of pollution

However, the air quality challenge is far from being solved. Although the number of people exposed to harmful air pollution has significantly decreased, persistent exceedances above World Health Organization guideline exposure levels remain for several air pollutants. Especially in the northern area of Italy known as Pianura Padana where Turin and Milan are among the most polluted cities in whole Europe.

For several years, in almost all European and North American cities, the urban air pollution levels characterization mainly relied on the direct measurements of the concentration of the different pollutant species.

Nevertheless, even if these measurements provide direct information of the air pollution level at given positions, they do not provide an exhaustive picture of the distribution of air pollution throughout urban areas nor the possibility to evaluate the impact of a new traffic plan. For this reason, during the last decade, public authorities have increasingly adopted pollutant dispersion models to complement the information provided by monitoring networks.

In the last decades, the team of LMFA (Laboratoire de Mécanique des Fluides et d'Acoustique) in the École Centrale de Lyon has been developing SIRANE, an urban air pollutant dispersion model conceived to simulate pollutant dispersion emitted from line sources (e.g. traffic emissions) and point sources (e.g. chimneys) at the district scale. This model has already been tested and validated against real data measurements. However, it requires a significant amount of computational time to simulate over a long time frame.

This thesis aims to develop statistical tools for predicting street-level pollutant concentrations under varying meteorological conditions. The objective is to create a fast, general-purpose model for rapid predictions.

This work primarily focuses on the SIRANE simulation software and its use in generating a statistically significant dataset of scenarios upon which street-level NO_x concentration can be inferred. An interpolation-based surrogate model and a regression-based predictor were developed to estimate pollutant concentrations from meteorological variables. The interpolation model achieved the highest overall accuracy, leveraging the data structure itself, while the regression approach offered a more generalizable and lightweight alternative. Although regression performance remained slightly lower, particularly in capturing local street-level variability, it proved effective in reconstructing temporal concentration patterns with limited input complexity.

Acknowledgments

I would like to express my deep gratitude for the opportunity I was given to take part in this thesis project, and for the support and guidance provided throughout the work, to my supervisors: Luca, Sofia, Pietro, and Chi Vuong. The help I received proved invaluable during the most challenging moments of the project.

I would like to thank all the members of the LMFA laboratory in Lyon for the warm welcome I was given. I am grateful for this experience, which helped me grow both personally and professionally, and allowed me to discover with great pleasure the world of research and academia.

I would like to thank my family, especially my mother, for always supporting and respecting my personal choices regarding my career and private life, for providing me with the right support — both moral and financial — and for offering constructive criticism when needed. I am happy to have such a wonderful safe shelter to support me.

I would like to thank my grandfather. I am sorry that I was not able to give you, in time, the joy you so often wished for. I hope at least to preserve, within these few lines, your strong spirit and your gentle wisdom, and to offer you a lasting memory by dedicating this work to you.

I would also like to thank my second family — my friends — with whom I have shared moments full of joy and carefree happiness, as well as times of harsh and painful reality. I am grateful to have had by my side people who have seen me grow and mature, and with whom I have learned to share the good times and overcome difficulties thanks to their companionship.

I would like to thank Margherita, for always standing by my side, never showing the slightest doubt in my abilities and worth, even when I had completely lost confidence in myself. Your presence has been, and will continue to be, precious to me. I am grateful for the moments of beauty you have given me; I am certain they will never leave me and will help me become a better person.

Finally, I would like to thank the HPC laboratory of Politecnico di Torino. Without their support, this work — which required considerable computational effort — would not have been possible. It is a precious resource and a source of pride for young students and researchers. Thank you for the professionalism and dedication in your work, which makes all of this possible.

Contents

List of Figures

List of Tables

Chapter 1

Urban Air Pollution

Urban air pollution represents one of the most pressing environmental challenges faced by modern cities.

This chapter provides an overview of the main sources of urban emissions, the physical processes that govern pollutant dispersion within the urban environment, and the associated impacts on human health and well-being. Understanding these fundamental processes sets the stage for introducing the modeling approaches discussed in the following chapters.

All the knowledge derives from online sources and the book Urban Climates [5].

1.1 Emissions and diffusion in the urban atmosphere

Urban areas are major focal points of pollutant emissions due to dense human activities. The dominant sources of urban air pollution today are anthropogenic, especially vehicle traffic. In Europe, the transport sector has overtaken industry and high-sulfur fuel combustion as the largest urban air pollution source. Road traffic contributes heavily to NO_x and particulate emissions in cities, despite technological advances like catalytic converters reducing per-vehicle emissions. Other significant urban emission sources include residential heating (particularly solid fuel or biomass burning in winter), industrial processes, construction activities (dust), and natural dust or pollen influx. These primary emissions consist of pollutants such as particulate matter (PM_{10} and $PM_{2.5}$), nitrogen oxides (NO_x), nitric oxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and volatile organic compounds (VOCs). Secondary pollutants like ozone (O_3) form within the urban

atmosphere from chemical reactions of precursors (NO_x and VOCs) under sunlight, contributing to urban smog.

Once emitted, pollutants in a city undergo dispersion (transport and diffusion) within the urban canopy and boundary layer. The physical layout of a city strongly influences this dispersion. Buildings and urban morphology create a complex roughness that slows wind and can trap pollutants at low levels. For example, a "street canyon" – a street flanked by buildings – acts as a semi-enclosed compartment that limits pollutant dispersal. Emissions from vehicles within street canyons tend to accumulate, making these locations pollution hotspots. Airflow in a street canyon often forms vortices that recirculate air; pollutants may remain trapped until enough wind penetrates the canyon to ventilate it. The dimensions and spacing of buildings (and trees) are significant: they determine mixing and ventilation efficiency in the urban canopy. Dense, tall building arrangements with narrow streets inhibit horizontal and vertical mixing, leading to higher concentrations near the source, whereas more open layouts allow pollutants to dilute more readily.

On the neighborhood to city scale, urban areas modify the atmospheric boundary layer, affecting how pollutants diffuse vertically. The layer of air just above the buildings (the urban canopy layer) extends up to mean building height, above which lies the urban boundary layer that can reach 1–2 km in height by day but contracts at night. During daytime, solar heating creates convective turbulence that promotes vertical mixing of pollutants. A deep mixed layer can dilute pollutants, carrying some of the burden upward away from the surface. At night, especially under clear, calm conditions, radiative cooling of the surface can lead to a temperature inversion (cool air trapped under warmer air aloft). This suppresses upward mixing, sometimes confining pollutants to a shallow layer near the ground. In urban areas, the result is often an overnight accumulation of pollution—a phenomenon sometimes described as an urban pollution dome when winds are light. Under these stagnant conditions, the city's emissions build up over the urban area in a quasi-circular "dome" of polluted air. This pollution dome can be reinforced by the urban heat island: the warmth of the city relative to its surroundings can induce circulations that recirculate pollutants. Indeed, a strong urban heat island (UHI) not only creates a thermal dome but also tends to self-sustain a pollution dome by trapping emissions in a recirculating flow, thereby prolonging smog episodes. In contrast, when regional winds are strong, the polluted air is advected downwind, forming an urban plume. The city then behaves like a giant "chimney," and pollutants are transported away in a plume that gradually disperses as it travels. Downwind communities may experience this transported pollution plume even as upwind areas remain cleaner.

Dispersion in urban environments is thus a balance between pollutant release rates and the capacity of the atmosphere to dilute and transport those pollutants. Atmospheric turbulence (enhanced by daytime heating and by air moving around buildings) acts to diffuse pollutants, whereas stable stratification and weak winds allow concentrations to build. Furthermore, chemical transformation plays a role: for example, nitric oxide (NO) emitted by vehicles can react with ozone at night, depleting local ozone – but this process is limited to the shallow layer of the nocturnal city air, underscoring the importance of vertical mixing between the surface and layers aloft.

Another critical aspect of urban diffusion is the interplay of pollution with climate factors. In humid atmospheres, pollutants can act as condensation nuclei, contributing to haze or smog formation. Historically, coal-burning cities suffered from industrial smog (so-called "pea-soup fogs"), where smoke and sulfur dioxide mixed with fog. A notorious example is the London smog of 1952, during which cold, stagnant air led to extreme accumulation of smoke; roughly 4,000 people died during the week of the smog, and total excess deaths reached an estimated 12,000 in the aftermath. Such events were exacerbated by temperature inversions trapping pollutants. Modern cities, having reduced coal smoke, now more commonly face photochemical smog. This type of smog, characterized by a brownish haze, results from sunlight-driven reactions of NO_x and VOCs (mainly from vehicle exhaust). Photochemical smog is typically a daytime, summer phenomenon producing irritants like ozone. Episodes have been observed in cities worldwide – for instance, in Los Angeles and also in European cities on sunny days with heavy traffic. In these conditions, the combination of intense sun and stagnant air leads to elevated ozone and fine particles, causing reduced visibility and acute health symptoms (e.g. eye irritation, coughing). The brown haze often seen hanging over large cities is an indication of such photochemical pollution, which can trigger breathing difficulties and asthma attacks in the populace.

In summary, emissions in urban environments originate largely from human activities (vehicles, heating, industry) and their dispersion is governed by urban microclimates. Complex airflow around buildings, atmospheric stability, and urban heat island effects all modulate how pollution spreads. A city can alternate between trapping its pollutants in a dome under static weather or flushing them out in plumes under ventilated conditions. Understanding these processes is essential, as it underpins both pollution forecasting in cities and the design of mitigation strategies (for instance, urban designs that promote ventilation, or timing industrial emissions to favorable dispersion periods). Ultimately, the goal is to avoid the worst-case scenario of stagnant, high-pollution episodes, and instead ensure that natural diffusion

and transport processes keep urban air quality within safe limits.

1.2 Impact on health and wellness

Air pollution is widely recognized as one of the most severe environmental health risks. The inhalation of polluted urban air has both immediate and long-term effects on human health and well-being. In the short term, high pollution episodes (e.g. acute smog events) lead to increases in hospital admissions for respiratory and cardiovascular distress, trigger asthma attacks, and can even cause acute poisoning or death in extreme cases. Over the long term, chronic exposure to polluted air contributes to the development of diseases and higher mortality rates in the population. The World Health Organization estimates that ambient (outdoor) air pollution is linked to 4.2 million premature deaths globally per year, primarily due to heart disease, stroke, chronic obstructive pulmonary disease (COPD), lung cancer, and acute respiratory infections. This makes air pollution a leading cause of death worldwide, on par with major health risk factors.

In urban areas, where pollution levels are highest, the public health burden is especially pronounced. A recent analysis in Europe highlighted that air pollution is the continent's single biggest environmental health risk, implicated in over 240,000 premature deaths each year in Europe. Prolonged exposure to fine particulate matter ($PM_{2.5}$) and other pollutants has been linked not only to respiratory illnesses and cardiovascular diseases, but also to outcomes like lung cancer. There is emerging evidence of links to other conditions such as diabetes and adverse birth outcomes, underlining that pollution's impact on health is systemic. The European Environment Agency (EEA) and WHO have noted that certain groups – children, the elderly, and people with pre-existing conditions like asthma – are particularly vulnerable. For example, children growing up in polluted cities may suffer from reduced lung development and more frequent bronchitis and asthma symptoms. The elderly and those with heart or lung diseases are at higher risk of hospitalizations or death when pollution spikes, as their systems are less able to cope with the added stress on lungs and blood circulation.

The specific pollutants in urban air each carry their own health risks. Particulate matter is especially concerning: $PM_{2.5}$ (particles $\leq 2.5\mu m$) can penetrate deep into the lungs and even enter the bloodstream, contributing to inflammation that affects the heart and other organs. Long-term $PM_{2.5}$ exposure is strongly associated with higher risks of heart attacks, strokes, lung cancer, and reduced life expectancy. Coarser PM_{10} ($\leq 10\mu m$) primarily affects the upper airways and can lead to respiratory irritation, coughs, and aggravation of asthma or chronic bronchitis. Epi-

demiological studies have consistently shown that for every incremental increase in fine particulate pollution, there is a measurable increase in mortality and morbidity. In fact, life expectancy in heavily polluted urban regions is years shorter than in cleaner areas, largely due to pollution-driven health effects. Nitric oxide (NO), nitrogen dioxide (NO_2), and the broader group nitrogen oxides (NO_x) are also key pollutants. NO is a primary product of combustion, particularly from vehicles, and rapidly oxidizes to NO_2 in the atmosphere. NO_2 is a potent respiratory irritant that can inflame the lungs and impair immune responses, especially in children. These pollutants also serve as precursors to ozone and fine particulate formation, compounding their health impacts. Chronic NO_2 exposure inflames the lining of the lungs and can reduce immune response, leading to more frequent infections (especially in children). NO_2 also serves as a proxy indicator for traffic-related pollution, which includes a mix of toxic constituents. Ozone (O_3) at ground level (a secondary pollutant) causes irritation of the eyes, nose, and throat and can trigger asthma attacks; high ozone days often see upticks in emergency room visits for breathing problems. Sulfur dioxide (SO_2), more common historically (from coal burning), can cause bronchoconstriction and coughing even at low concentrations in sensitive individuals. Though SO_2 levels have declined in many cities due to cleaner fuels, it remains a concern in regions still reliant on coal or heavy oil. Carbon monoxide (CO) from incomplete combustion can acutely affect oxygen delivery in the body (by binding to hemoglobin), and high concentrations (often in enclosed or poorly ventilated spaces like tunnels or garages) can be immediately dangerous; in open city streets, CO generally stays below acutely toxic levels but contributes to chronic low-level exposure.

Beyond these physical health impacts, urban pollution also degrades overall quality of life or "wellness." On smoggy days, people may be advised to stay indoors, reducing opportunities for exercise and recreation, and causing psychological stress or anxiety about health. The visibility reduction and noxious odors of smog can also cause discomfort and mental distress. There is growing research into links between chronic pollution exposure and mental health or cognitive development – for instance, associations with higher risk of neurodevelopmental issues in children and neurodegenerative diseases in the elderly, although these links are still being studied. What is clear is that cleaner air correlates with a healthier, more productive population: fewer missed work or school days, lower healthcare costs, and better overall well-being.

From an epidemiological perspective, the burden of disease attributable to urban air pollution is immense. A European study found that in some of the most polluted cities, life-long residents lose multiple years of life expectancy compared to those in

cleaner cities. Children suffer increased rates of asthma and bronchitis; the World Health Organization reported that even in relatively developed regions, thousands of deaths per year in children and adolescents can be attributed to air pollution. These stark statistics have framed air pollution as not just an environmental issue but a public health crisis. Indeed, air pollution's impact is sometimes compared to that of smoking. Recent global assessments suggest that breathing chronically polluted air may be as harmful as smoking several cigarettes per day in terms of health risk, and globally more people now die from air pollution than from tobacco each year.

In conclusion, urban air pollution markedly undermines health and wellness. It increases acute illness and mortality during high-pollution episodes and contributes silently to chronic disease and premature death over time. The most dangerous pollutants – fine particulates and ozone – penetrate deep into the body, causing inflammation and oxidative stress that affects organs far beyond the lungs. Efforts to improve urban air quality are therefore critical public health interventions. Every reduction in pollutant concentrations can yield measurable improvements in population health, from better respiratory function in children to fewer heart attacks in adults. This is a major motivator behind environmental regulations and pollution control policies, as discussed in the next section.

1.3 Enviromental policies

Given the significant impacts of urban air pollution on health and the environment, governments have developed a variety of policies and measures to reduce emissions and improve air quality. Historically, some of the earliest and most dramatic clean air policies were triggered by acute pollution disasters. A landmark example is the response to the London Great Smog of 1952. In its wake, the UK passed the Clean Air Act of 1956, which restricted the burning of coal in urban areas, encouraged a shift to cleaner fuels, and mandated taller chimney stacks for industry to disperse emissions higher into the atmosphere. These measures successfully eliminated the dense, deadly smoke-fogs ("smogs") that once frequently shrouded London and other industrial cities. Similar legislation followed in other countries (for example, the United States Clean Air Act of 1970 and its amendments) targeting industrial emissions, vehicle exhaust, and fuel quality.

Over the decades, a robust framework of air quality regulation has evolved at international, national, and local levels. Air quality standards form the backbone: scientific research (often synthesized by the WHO) informs guideline concentration limits for key pollutants, which governments then adopt as legal standards. For

instance, the European Union's ambient air quality directives set limit values for pollutants like PM_{10} ($50\mu g/m^3$ daily, not to be exceeded more than 35 times per year) and NO_2 ($40\mu g/m^3$ annual mean) to protect public health. Cities and countries are required to monitor air quality and report if these standards are violated, and to implement Air Quality Plans to achieve compliance. Over time, the EU has periodically tightened these limits (and is in the process of revising them further to approach WHO's recently updated guidelines). In practice, these standards have driven many policy actions "on the ground."

One major policy approach has been emission control at the source. For vehicles, this includes progressively stricter exhaust emission standards (Euro 1 through Euro 6/VI in Europe, Tier standards in the US, etc.) that mandate cleaner engines and the use of pollution control devices like catalytic converters and particulate filters. As noted, the introduction of catalytic converters and the phase-out of leaded petrol in Europe led to substantially lower emissions per car. However, the growing number of vehicles meant that without further measures, overall pollution could still rise. Thus, cities have also implemented traffic management and vehicle restrictions. Low Emission Zones (LEZs) have been established in hundreds of European cities – these zones allow only vehicles meeting certain emission criteria to enter, effectively banning older, high-emitting cars from city centers. For example, cities like London, Berlin, and Paris have LEZ or even Ultra Low Emission Zone schemes targeting diesel soot and NOx. Some cities have experimented with congestion charging (as in London or Stockholm) which indirectly improves air quality by reducing traffic volumes. Others, especially in developing countries, have adopted alternate license plate driving days or outright driving bans during smog episodes.

Industrial emissions are addressed through regulations on factories and power plants – requiring smokestack scrubbers, filters, and sometimes relocation of heavy industries away from densely populated areas. After mid-20th century policies mandated taller chimneys and emission controls, the infamous industrial smogs dissipated. Today, large combustion plants in many countries must adhere to strict emission limits for SO_2, NO_x, and particulates, and they are often part of cap-and-trade schemes for pollutants or greenhouse gases. Fuel quality improvements have also been critical: removing sulfur from fuels (to near-zero levels in gasoline and diesel by the 2000s in the EU and U.S.) has cut SO_2 emissions drastically and also allowed cleaner vehicle technologies. In many cities, transitioning power generation and heating from coal to cleaner alternatives (natural gas, renewables) has yielded major air quality gains, as seen in European cities from the 1980s onward.

At the urban planning level, cities are increasingly incorporating air quality considerations. Urban design can improve ventilation – for instance, preserving open park

spaces and breezeways, and avoiding the creation of too many deep street canyons in new developments. "Green infrastructure" like urban trees and green roofs can help by not only improving aesthetics and reducing urban heat (which can exacerbate ozone formation) but also by capturing some particulate matter on leaves (though this is a modest effect relative to emissions reductions). Some cities promote these as part of climate adaptation and air quality improvement strategies. Initiatives to expand public transportation, cycling, and walking infrastructure also serve to reduce reliance on high-pollution modes of transport.

Another critical policy area is the episode (emergency) response plan. Many urban regions have tiered alert systems that trigger short-term measures when pollution is forecasted to spike. For example, authorities may impose temporary driving bans or speed reductions, encourage car-pooling and remote work, reduce industrial production, or ban wood burning in fireplaces during multi-day pollution episodes. These are often communicated as pollution "alert levels" (e.g., Code Red days). While such emergency measures are stop-gaps and not substitutes for long-term solutions, they can mitigate the worst peaks. Public awareness campaigns are also a policy tool – educating citizens on actions to reduce emissions (like not idling cars, or using public transit on bad air days) and on how to protect themselves (such as avoiding outdoor exercise when air quality is poor).

Internationally, cooperation has helped drive policies. The transboundary nature of air pollution (e.g., pollutants can travel long distances downwind) led to agreements like the UNECE Convention on Long-Range Transboundary Air Pollution (CLRTAP) in 1979. Under CLRTAP, protocols were established to cut emissions of SO_2, NO_x, VOCs, and other pollutants across Europe and North America. The EU's multi-pollutant directives (National Emission Ceilings Directive) and programs like the Auto-Oil initiatives in the 1990s were born from recognition that comprehensive strategies were needed. These set national caps on total emissions and promoted cleaner vehicle technology and fuels. More recently, concerns about climate change have dovetailed with air quality policy: cutting CO_2 often coincides with reducing co-emitted pollutants (for instance, shifting from coal to renewable energy reduces CO_2 and also SO_2/PM). Thus, climate policies (like the EU Green Deal or city climate action plans) often have co-benefits for air quality.

The effectiveness of these policies is evident in many success stories. Across most high-income countries, lead, sulfur dioxide, and carbon monoxide levels have plummeted since the 1970s due to fuel and industrial regulations. In the EU, average urban PM_{10} levels have declined and the number of days with extreme pollution has generally fallen in the last few decades. However, challenges remain. NO_2 and $PM_{2.5}$ levels in many cities still regularly exceed health-based limits, and enforce-

ment of regulations can be inconsistent. Moreover, in rapidly developing cities in Asia and Africa, pollution levels are still extremely high, calling for aggressive policy interventions similar to – or beyond – those implemented in Europe or North America.

In conclusion, environmental policies addressing urban air pollution span a wide range: from technological standards and fuel regulations to traffic restrictions and urban planning. A combination of these strategies is usually needed to tackle the complex mix of sources in a city. Crucially, policy development is an ongoing process: as scientific understanding of health effects improves (often revealing harm at lower concentrations than previously thought), governments are prompted to tighten standards and innovate new solutions. Urban air pollution control stands as a testament to how policy, technology, and public awareness must work together to produce healthier city environments.

1.4 The air pollution in Turin

Turin (Torino), a major city in northern Italy, offers a case study in urban air pollution, illustrating both the severity of the problem and the efforts to mitigate it. Turin lies in the Po Valley, a region notorious for having some of the highest pollution levels in Europe. The geography and climate of the Po Valley strongly predispose it to pollution accumulation. The valley is bordered by the Alps to the west and north of Turin and by the Apennine mountains to the south, creating a large basin with relatively weak natural air circulation. Especially in winter, temperature inversions are common: cold air pools in the valley under a lid of warmer air, effectively turning the Po Valley into a bowl that traps pollutants. As a result, emissions from vehicles, industry, and heating across Turin and other Po Valley cities tend to build up rather than disperse. The region often experiences extended periods of stagnant air, during which pollutant concentrations can surge to very high levels. In essence, the Po Valley acts as a containment area where both primary pollutants (like PM and NO_2) and secondary pollutants (like ozone formed from precursors) accumulate, particularly during late autumn and winter when ventilation is poorest.

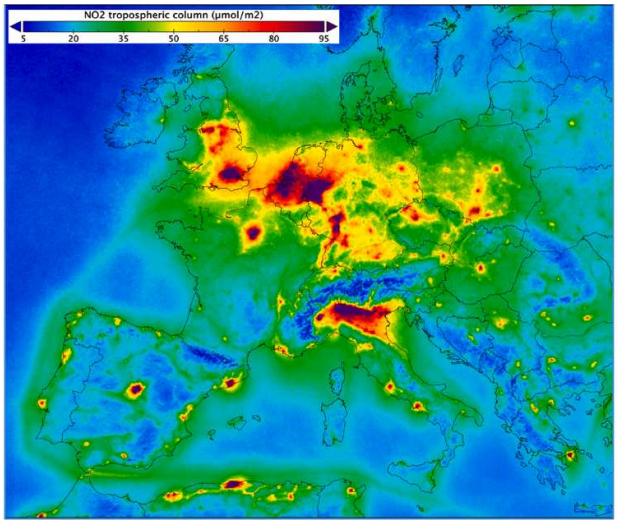


Figure 1.1: NO2 EU concentration from ESA Copernicus Sentinel-5P (ref period 04–2018 to 03–2019) from [2]

The Alps (top) and Apennines (bottom) surrounding the valley form a barrier that inhibits pollutant dispersion. Turin's urban emissions profile is similar to that of other large cities, with transportation, residential heating, and industry being key contributors. However, its location in the heart of the Po Basin amplifies the impact of these emissions on air quality. Measurements consistently show Turin to be among the most polluted cities in Italy. For example, in 2022 Turin had the highest number of days with PM_{10} levels exceeding the EU daily limit of $50\mu g/m^3$. One of the city's monitoring stations (Torino-Grassi) recorded 98 days over the limit in 2022 – nearly triple the 35 exceedances per year permitted by law. This was the worst in the country, ahead of other Po Valley cities like Milan (84 days) and others in the same year. Such statistics underscore how frequently Turin's residents are breathing unhealthy air. Furthermore, Turin's annual average concentrations of pollutants are well above both EU standards and WHO guidelines. The annual mean PM_{10} in Turin in 2022 was around $37\mu g/m^3$, nearly double the current EU annual limit ($20\mu g/m^3$ from 2030, $40\mu g/m^3$ previously) and far above the WHO recommended $15\mu g/m^3$ (and the even lower $5\mu g/m^3$ guideline updated in 2021 for $PM_{2.5}$). For finer particles, $PM_{2.5}$, the levels in Turin (often in the $20 - 25\mu g/m^3$ range annually) exceed the WHO guideline of $5\mu g/m^3$ by about 4 to 5 times. A recent analysis noted that residents across the Po Valley routinely breathe air with particulate concentrations several times higher than what WHO deems safe. This translates into significant health impacts locally – elevated rates of asthma and other respiratory illnesses, and a contribution to cardiovascular disease and mortality. Indeed, Italy sees tens of thousands of premature deaths each year attributable to air pollution, and the

highest per-capita toll is in the northern regions encompassing Turin and the Po Valley.

The pollutant mix in Turin includes high levels of NO_2, largely from diesel traffic. In 2019, for instance, Turin and Milan had some of the highest NO_2 averages in Europe, often exceeding $50 - 60\mu g/m^3$ in annual mean (versus the EU limit of $40\mu g/m^3$ and WHO guideline of $10\mu g/m^3$). In recent years, there have been incremental improvements – partly due to newer vehicle fleets and temporary reductions during events like COVID lockdowns – but NO_2 pollution remains a serious concern. Ozone (O_3) in summer also affects Turin; the sunny climate and ample precursor emissions in the region lead to frequent breaches of ozone target values in the warmer months, especially downwind of the city. In winter, however, particulate matter is the headline pollutant. Turin's cold season PM problems stem from a combination of vehicle exhaust, domestic heating (some residents in the region still use wood or pellet stoves), and industrial and agricultural emissions (including ammonia from agriculture contributing to secondary PM formation). The stable meteorological conditions from roughly October through March mean that Turin often starts accumulating pollution in early winter and does not get sustained relief until spring winds or rains arrive.

Local and regional authorities are acutely aware of these challenges, and Turin has become a focal point for pollution-control initiatives in Italy. Piedmont (the region containing Turin) together with neighboring regions in the Po Valley has adopted a coordinated air quality plan – indeed, the Piano Regionale Qualità dell'Aria was updated in 2018 to align with the broader Po Basin agreement. Under this plan, a suite of measures are implemented in Turin:

- Seasonal vehicle restrictions: Every year during the cold months (typically October 1 to March 31), Turin enforces progressive bans on older vehicles. For example, petrol cars below Euro 2 and diesel cars below Euro 3 or Euro 4 are not allowed to circulate on weekdays. These restrictions tighten under an agreed protocol if pollution persists. Notably, when pollution levels reach defined thresholds for consecutive days, emergency traffic bans are triggered. Turin has a tiered alert system (often color-coded from green to red). At the highest alert (e.g. "red alert"), even relatively modern diesels (Euro 5 and below) can be temporarily banned from city roads. In late 2017, after successive days of high PM, Turin raised its smog alert to red and banned all diesel cars up to Euro 5 during daylight hours, sidelining about half a million vehicles. More recently, in February 2024, amid another pollution episode, Piedmont authorities imposed a temporary ban on diesel vehicles Euro 3 through Euro 5 in Turin and surrounding towns for several days. Such

measures underscore the drastic steps taken to curb emissions when air quality degrades.

- Heating and industry measures: During high-pollution alerts, Turin also restricts residential and commercial heating—requiring lower thermostat settings (e.g. max 19°C in buildings) and urging a switch from wood-burning to cleaner heating if possible. The local government has incentivized the phase-out of old wood-burning stoves and diesel generators. Industrial facilities in the area are required to curtail activities or use extra emission abatement on bad air days. The city is also expanding its district heating network, which allows centralized, cleaner heat production (from cogeneration plants) instead of numerous individual boilers.

- Urban transport and planning: Turin has invested in public transport (with a modern metro line and plans for expansion, plus promotion of electric buses) and bike-sharing infrastructure in an effort to provide alternatives to car use. The city center has a Limited Traffic Zone (ZTL) in which vehicular access is restricted or requires special permits, aiming to reduce congestion and pollution in the dense urban core. Additionally, Turin is one of the cities participating in the EU funded LIFE Prepair project: a collaborative initiative among regions in the Po Valley to implement and share best practices for emissions reduction (covering areas like vehicle emissions, energy efficiency, and agricultural emissions). Under such programs, Turin has been piloting low-emission vehicle incentives, installing more electric vehicle charging stations, and testing innovative solutions (for example, using photocatalytic materials on some road surfaces that might help break down pollutants).

Despite these efforts, progress is incremental. Data show that while there have been improvements over the past two decades (for instance, average PM_{10} levels in Turin have come down compared to the early 2000s, and the number of annual mega-smog episodes has slightly decreased), Turin's air quality still regularly fails to meet both EU legal standards and the ambitious WHO health guidelines. For instance, in 2023, Turin was again among the cities with the highest number of PM_{10} exceedance days (though slightly fewer than 2022, thanks in part to favorable weather). Legambiente's annual report Mal'Aria continually places Turin at or near the top of Italy's pollution rankings, a stark indicator of the work remaining. Local authorities acknowledge that without substantial reduction in emissions at source – especially a faster renewal of the vehicle fleet to low-emission and electric vehicles, and more fundamental shifts such as moving freight off roads – the city will struggle to attain compliance with upcoming stricter EU air standards (like the new $PM_{2.5}$

limits for 2030). In fact, analyses suggest Turin (along with Milan) will need roughly a 43% reduction in PM_{10} emissions to meet the 2030 targets, one of the steepest cuts required among Italian cities.

On a positive note, awareness among citizens is high, and there is increasing public pressure for cleaner air in Turin. Grassroots movements (e.g. parents campaigning for clean air around schools) have emerged, and the issue of smog is a regular topic in local media each winter. The municipality has thus integrated air quality into its broader sustainability plans. The Torino 2030 Sustainable and Resilient strategy explicitly includes an Air Quality Plan and a push for green infrastructure and climate resiliency, recognizing that mitigating heat islands and increasing green spaces can complement air pollution reduction. Moreover, Turin has joined international networks of cities (like C40 Cities and Eurocities) to exchange knowledge on combating pollution and climate change. These collaborations help Turin implement measures that have worked elsewhere, such as zero-emission zones or pollution pricing schemes, potentially in the future. In summary, Turin exemplifies the plight of a city facing significant air pollution due to both human factors and geographic destiny. Located in the pollutant-trapping Po Valley, it endures frequent smog episodes especially in winter, with pollutant levels that challenge health limits. The city and region have responded with a multipronged strategy: emergency traffic bans during peak smog, progressive elimination of older polluting vehicles, improvements in heating and industrial emissions, and long-term urban planning for sustainability. While these measures have prevented conditions from becoming as dire as in the mid-20th century, Turin's air remains far from clean. Achieving sustained, substantial improvements will likely require continued policy innovation, technological change (such as a transition to electric mobility), and persistent public and political will. The case of Turin underscores that solving urban air pollution is a long-term endeavor – one that must adapt to new scientific findings and balance economic, social, and environmental considerations – but it is an endeavor critical to the health and quality of life of the city's residents.

Chapter 2

Modelling Urban Air Pollution

In order to control and manage urban air quality, public authorities require an integrated approach that incorporates direct measurements and modelling of mean pollutant concentrations. These have to be performed by means of operational modelling tools, that simulate the transport of pollutants within and above the urban canopy over a large number of streets. The operational models must be able to assess rapidly a large variety of situations and with limited computing resources.

This chapter describes the operational modeling tools available for simulating urban air pollution, with emphasis on the SIRANE model. After presenting the general principles of air quality measurement and computational modeling, the chapter provides a detailed technical overview of the SIRANE framework and its input requirements, concluding with validation results from a real-world case study.

2.1 How is pollution measured

2.1.1 Monitoring stations

Air pollution in urban areas is commonly measured using a network of monitoring stations, typically managed by regional environmental protection agencies. These stations are strategically positioned to capture variations in pollutant concentrations across different urban settings: traffic-heavy roads, industrial areas, residential zones, and background rural locations.
Modern monitoring stations are equipped with automatic analyzers that provide high-resolution time series (e.g., hourly averages) of pollutants such as NO, NO_2, NO_x, O_3, CO, SO_2, PM_{10}, and $PM_{2.5}$. These stations often include meteorological sensors (wind speed and direction, temperature, humidity, and solar radiation),

which are crucial for interpreting air quality data and modeling pollutant dispersion. Passive samplers are also employed to capture average concentrations over longer periods (typically weeks), especially in areas without fixed stations. These are useful in dense networks for spatial mapping, although they provide limited temporal resolution.

2.1.2 Modelling tools

Predictive models are essential complements to monitoring networks. They provide spatial and temporal coverage beyond what physical stations can offer and are crucial for scenario analysis and urban planning. Most modelling tools are divided into two different categories based on the computational model that is used in order to predict the pollutants' concentration level:

- **CFD codes**: Computational Fluid Dynamics (CFD) models numerically solve the Navier-Stokes equations to simulate airflow and pollutant dispersion in high resolution. They allow for detailed simulations of airflow in complex urban geometries but are computationally expensive. Examples include OpenFOAM-based models or the PALM model system.

  Despite their accuracy, CFD codes are generally limited to small-scale applications or specific domains (e.g., a few city blocks) due to computational demands. They are valuable for studying detailed dispersion in specific settings like street canyons or intersections.

- **Parametrization of mass and momentum transfer**: For larger-scale and operational models, simplified representations of fluid dynamics are used. These parametrizations rely on empirical or semi-empirical formulations to model:

  1. **Vertical exchanges** between the urban canopy and the boundary layer
  2. **Street-level transport** driven by convective and turbulent mechanisms
  3. **Lateral and vertical dispersion** based on Monin-Obukhov similarity theory and Gaussian plume models

  These approximations allow real-time or near-real-time predictions at city scale and form the basis of urban-scale models like SIRANE.

.

2.2 SIRANE

2.2.1 Model outline

The SIRANE model (**Soulhac et al., 2011**[10]) is a deterministic, operational urban dispersion model designed to estimate mean pollutant concentrations at the district scale, accounting for the complex morphology of urban street networks. It operates under a steady-state approximation over hourly time steps and is specifically adapted for the simulation of emissions from line sources (e.g., traffic) in dense urban environments. The model decomposes the urban domain into two main regions: the *urban canopy layer* (i.e., the volume inside street canyons) and the *overlying atmospheric boundary layer*.

This work focuses exclusively on the estimation of pollutant concentrations at street level. Consequently, the description of Gaussian atmospheric dispersion processes above roof level is omitted. The analysis is restricted to the mechanisms that govern pollutant accumulation and exchange within the street network.

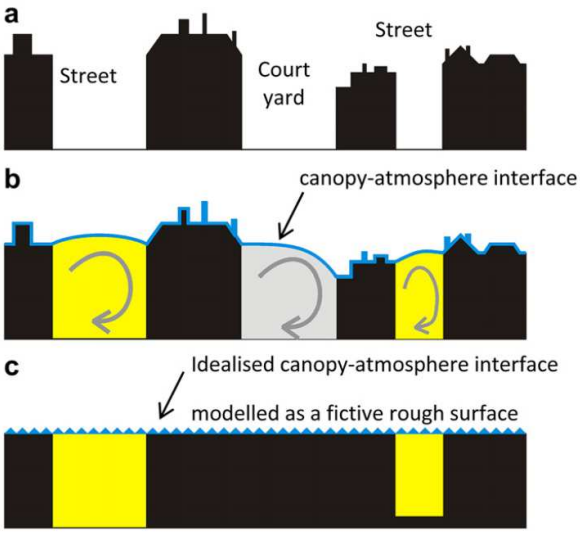


Figure 2.1: Description of the urban domain in the model SIRANE.

Structure of the domain:

The urban district is represented as a network of connected street segments and intersections (Fig: 2.2a). Each street is modeled as a rectangular channel (a "box") characterized by its length L, width W, and height H (average building height). Pollutant concentrations are assumed to be uniformly mixed within the street volume (well-mixed box hypothesis). Pollutants can exit the canyon through three main processes:

- Convective mass transfer along the street due to the mean wind along their axis (Fig: 2.2b);
- Turbulent transfer across the interface between the street and the overlying atmospheric boundary layer(Fig: 2.2b);
- Convective transport at street intersections. Exchange ratios have been parameterised by a mass exchange tensor, which accounts for the influence of the external flow direction(Fig: 2.2a);

These three processes, along with the equations that guides them, are described in the following pages.

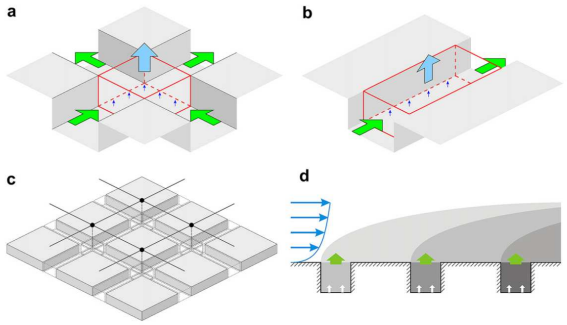


Figure 2.2: The different components of the SIRANE model. (a) Modelling a district by a network of streets. (b) Box model for each street, with corresponding flux balance (c) Fluxes at a street intersection. (d) Modified ed Gaussian plume for roof level transport.

Mass transfer along the street canyon:
Assuming steady state conditions, this mass balance over the street volume can be written as:

$$Q_S + Q_I + Q_{\text{part},H} = Q_{H,\text{turb}} + HWU_{\text{street}}C_{\text{street}} + Q_{\text{part},gr} + Q_{\text{wash}} \tag{2.1}$$

where:

- Q_S is the mass flux of pollutants emitted by sources within the street (e.g., vehicle emissions) $[\mu g/s]$;
- Q_I is the advective flux of pollutants entering the street from upstream via the street axis $[\mu g/s]$;
- $Q_{\text{part},H}$ is the sedimentation flux of solid particles entering the street from above through the street-atmosphere interface $[\mu g/s]$;

- $Q_{H,\text{turb}}$ is the vertical turbulent flux of pollutants from the street to the atmosphere (roof-level exchange) $[\mu g/s]$;
- $HWU_{\text{street}}C_{\text{street}}$ is the advective flux of pollutants exiting the street downstream along the street axis, where:
  - H is the average building height $[m]$,
  - W is the street width $[m]$,
  - U_{street} the spatially averaged wind velocity along the street axis $[m/s]$,
  - C_{street} is the spatially average concentration in the street $[\mu g/m^3]$.
- $Q_{\text{part},gr}$ is the deposition flux of solid particles toward the ground $[\mu g/s]$;
- Q_{wash} is the wet deposition flux due to precipitation scavenging $[\mu g/s]$.

This equation represents a balance between the incoming pollutant fluxes (emissions, upstream advection, particle sedimentation) and outgoing fluxes (turbulent ventilation, downstream advection, ground deposition, and wash-out). The model operates under a steady-state assumption for each hourly time step.
Each term is then expressed as a function of local variables and parameterized according to urban geometry and meteorology.

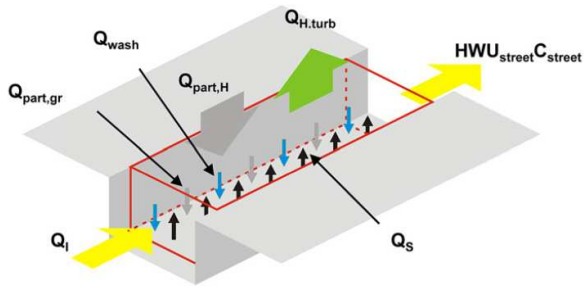


Figure 2.3: Mass balance within a street canyon.

(i) Advection along the street axis
Advection is driven by the spatially averaged wind velocity U_{street} parallel to the street axis. The advective flux through a cross-section $A = H \cdot W$ is given by:

$$Q_{\text{H,turb}} = AC_{\text{street}}U_{\text{street}} = HWU_{\text{street}}C_{\text{street}} \tag{2.2}$$

In a network of streets, each street receives inflow from upstream segments and contributes to downstream ones. The direction and magnitude of U_{street} are computed from the external wind speed and direction, adjusted for street orientation

and urban roughness. According to Soulhac et al. (2008) [8] U_{street} can be written as:

$$U_{\text{street}} = U_H \cos(\varphi) \frac{\delta_i^2}{HW} \left[\frac{2\sqrt{2}}{C}(1-\beta)\left(1 - \frac{C^2}{3} + \frac{C^4}{45}\right) + \beta \frac{2\alpha - 3}{\alpha} + \left(\frac{W}{\delta_i} - 2\right) \frac{\alpha - 1}{\alpha} \right] \tag{2.3}$$

The parameters used in this expression are defined as follows:

- U_H: mean wind velocity at the top of the internal boundary layer, computed from friction velocity u_* and stability functions (see below).
- φ: angle between the wind direction and the street axis [rad].
- H: average building height [m].
- W: street width [m].
- δ_i: thickness of the internal boundary layer within the canyon. Defined as:

$$\delta_i = \min(H, W/2)$$

- $\alpha = \ln\left(\frac{\delta_i}{z_{0,\text{build}}}\right)$: logarithmic stability parameter based on the building roughness length.
- $\beta = \exp\left[\frac{C}{\sqrt{2}}\left(1 - \frac{H}{\delta_i}\right)\right]$: stability correction factor.
- $z_{0,\text{build}}$: roughness length characterizing the urban canopy (typically a fraction of the building height, e.g. $z_0 \approx 0.1H$).
- C: dimensionless stability-dependent parameter, obtained by solving the following transcendental equation:

$$\frac{z_{0,\text{build}}}{\delta_i} = \frac{2}{C} \exp\left[\frac{\pi}{2}\frac{Y_1(C)}{J_1(C)} - \gamma\right]$$

  where:

  - $J_0(C), J_1(C)$: Bessel functions of the first kind;
  - $Y_0(C), Y_1(C)$: Bessel functions of the second kind;
  - γ: Euler–Mascheroni constant (≈ 0.5772).

- U_H is further given by:

$$U_H = u_* \sqrt{\frac{\pi}{\sqrt{2}k^2 C} \left[\frac{Y_0(C) - \frac{J_0(C)Y_1(C)}{J_1(C)}}{Y_1(C)} \right]}$$

  where:

  - u_*: friction velocity [m/s], representing surface shear stress;
  - k: von Kármán constant (typically $k \approx 0.4$).

This formulation accounts for the modification of the wind field due to the urban canopy's roughness and geometry. The use of Bessel functions arises from analytical solutions to the diffusion equations in the roughness sublayer above the urban canopy (Soulhac et al., 2011 [10]).
The term $\cos(\varphi)$ ensures that only the component of the wind parallel to the street axis contributes to along-street advection. The result is a spatially variable wind field across the network, depending on orientation and morphology.

(ii) Exchange with the atmosphere at roof level
Vertical mass exchange between a street canyon and the over- lying atmospheric flow is modeled using a parametrized exchange velocity u_d. The flux of pollutants exchanged by turbulent diffusion at the street- atmosphere interface is:

$$Q_{\mathrm{H,turb}} = u_d W L (C_{\mathrm{street}} - C_{\mathrm{street,ext}}) \tag{2.4}$$

where:

- u_d is the mass exchange velocity between the street and the overlying layer $[m/s]$.
- $C_{\mathrm{street}} - C_{\mathrm{street,ext}}$ is the mean pollutant gradient of concentration between the street and the atmospheric layer above roof level.

The exchange coefficient u_d accounts for turbulent transfer driven by roof-level shear and thermal convection, and is estimated (Soulhac, 2000 [9]; Salizzoni et al., 2009a [6]) to be :

$$u_d = \frac{\sigma_w}{\sqrt{2}\pi} \tag{2.5}$$

where σ_w is the standard deviation of the vertical velocity computed at roof level.

(iii) Exchange at intersections
At junctions, pollutant mass is redistributed across intersecting streets. The upwind

pollutant flux $Q_{I,j}$ entering street j is a function of the exchange coefficient of P_{ij} which in turn depends on the wind direction φ.
However, this coefficient does not take into account the turbulent mixing but only the topology of the mean streamlines within the intersections (Soulhac et al., 2009 [6]). By assuming that the turbulent mixing within the intersection mainly depends on larger scale fluctuations, its effect can be modelled by considering the standard deviation of the wind direction. We can therefore define a time averaged exchange coefficient as:

$$\hat{P}_{ij}(\varphi_0) = \int f(\varphi - \varphi_0)\, P_{ij}(\varphi)\, d\varphi \tag{2.6}$$

with

$$f(\varphi - \varphi_0) = \frac{1}{\sigma_\varphi \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\varphi - \varphi_0}{\sigma_\varphi}\right)^2\right] \tag{2.7}$$

where φ_0 is the average wind direction and σ_φ is the standard deviation of the wind direction.
The flux entering the intersection from street j can be expressed by taking into account the contribution of all streets connected at the intersection:

$$Q_{Ij} = \sum_i \hat{P}_{ij}(\varphi_0) C_{\text{street,i}} + P_{\text{ext}\to j} C_{\text{I,ext}} \tag{2.8}$$

The second term of (2.8) represents the flux from the external flow (therefore related to a concentration $C_{\text{I,ext}}$) and which enters the intersection vertically. The flux $P_{\text{I}\to j}$ is computed considering only the case of air from the external flow entering the intersection ($P_{\text{vert}} < 0$) and assuming that the incoming flux is distributed in the streets downwind of the intersection proportionally to the flow rate in each street:

$$P_{\text{ext}\to j} = \max(-P_{\text{vert}}, 0) \frac{P_{\text{street},j}}{\sum\limits_{\text{street downwind to the intersection}} P_{\text{street},i}} \tag{2.9}$$

The SIRANE model provides a physically grounded, computationally efficient framework to simulate pollutant concentrations at the urban scale. By combining simplified fluid dynamics within the canopy with Gaussian dispersion above roof level (not adressed in this project), and integrating emission inventories and meteorological data, it enables city-wide pollution mapping with relatively modest computational effort.
However, it assumes quasi-stationarity over hourly intervals, and neglects memory

effects (i.e., pollutants emitted in previous hours are not retained). This implies reduced accuracy under stagnant conditions or during pollution build-up events. Furthermore, it models the street network as a connected system of idealized boxes, and does not resolve microscale turbulence. Nevertheless, its simplicity allows rapid simulations over large districts with realistic geometries, making it suitable for operational use and exposure mapping.

2.2.2 Input data

The SIRANE model requires several categories of input data to simulate the dispersion of pollutants at the urban scale. These include meteorological observations, a geometric description of the street network, emission inventories from multiple sources, and background pollutant concentrations.

Meteorological data

Meteorological input is critical for driving both the advection and the turbulent dispersion components of the model. In particular, the following parameters are needed on an hourly basis:

- Wind directionφ [°], measured at reference height (typically 10 m);
- Air temperature T [$^{\circ}C$];
- Friction velocity u_* [m/s], derived from Monin-Obukhov similarity theory; represents the shear stress at the surface and is a key parameter for turbulent momentum and mass exchange.
- Inverse Monin-Obukhov length L_{MO}^{-1} [m^{-1}], used to characterize atmospheric stability; its reciprocal (L_{MO}) represents the height above ground beyond which buoyancy effects dominate over mechanical turbulence, influencing vertical mixing and pollutant dispersion.

The external wind field is used to compute the along-street wind component U_{street} using the formulation presented in Eq. 2.3, which accounts for building geometry and roughness. The turbulent fluxes at the roof level, which govern vertical pollutant exchange, are derived from u_* and L_{MO} using parametrizations grounded in Monin–Obukhov theory.
In SIRANE applications such as the Lyon case study (Soulhac et al., 2017 [7]), meteorological data are obtained from a local synoptic station (e.g. Météo-France station in Bron). The statistics of the wind field (e.g., wind rose and stability

class distribution) are analyzed to assess the representativeness of the data over the simulation domain.

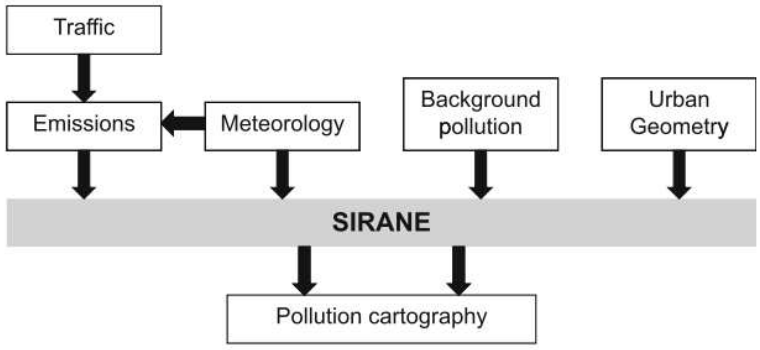


Figure 2.4: Scheme of SIRANE inputs and outputs from Soulhac et al., 2012 [11].

It is important to note that meteorological inputs in SIRANE can either be obtained from real measurements or can be synthetically generated. In sensitivity studies or surrogate model development, a synthetic dataset may be constructed by systematically varying meteorological parameters over plausible physical ranges. As will be shown in Section 4.2, an artificial meteorological dataset can be created by combining all possible values of: wind direction (φ), friction velocity (u_*), air temperature (T), inverse Monin-Obukhov length (L_{MO}^{-1}).
Such synthetic datasets enable parametric studies of pollutant dispersion in relation to atmospheric stability and wind conditions, and are particularly useful for training surrogate models or machine learning emulators based on SIRANE output.

The urban geometry: shape file

The urban geometry is described using a shapefile that defines the layout of the street network and associated urban canopy parameters. Each street is modeled as a polygonal segment characterized by:

- Street length L [m], width W [m], and average building height H [m];
- Classification as "canyon" or "open" based on the aspect ratio W/H;
  - Canyon $W/H \leq 3$;
  - Open: $W/H > 3$.
- Connectivity to other street segments via intersections;
- Geographic position in a projected coordinate system (EPSG:32632).

The shapefile is processed using GIS tools and pre-processing scripts (e.g., as described in Soulhac et al., 2011 [10]) to generate:

- A topological graph of the street network;
- A table of attributes per street (street ID, geometry);
- Identification of inflow/outflow segments for each intersection;
- Mapping between street IDs and emission inventory grid cells.

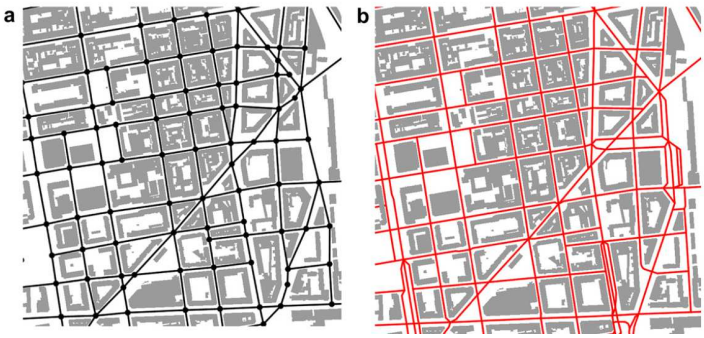


Figure 2.5: Example of how a shape file approximates the urban geometry.

Urban geometry encoded in the shape file is key in shaping pollutant dispersion patterns. It will play a central role in this thesis, as a reduced version of it—limited to the district of San Salvario—will be used throughout the analysis. All relevant details concerning its structure and preprocessing will be presented in Chapter 3.

Emission sources and temporal modulation

In the SIRANE modelling framework, emissions are categorized into four primary source types: line (or street) emissions, surface emissions, point emissions, and background concentrations. Each type plays a distinct role in shaping the resulting pollutant concentration field.

- **Linear emissions**: These refer primarily to road traffic and are associated with each segment of the street network. Traffic emissions represent the dominant source of NO and NO_2 in the urban canopy layer. Emission rates are spatially distributed along the road geometry and are modulated over time to reflect daily and weekly traffic patterns.
- **Surface emissions**: These cover distributed area sources such as domestic heating, commercial activities, or diffuse industrial emissions. They are

projected over polygons defined in the simulation domain and contribute to ground-level concentration fields, especially during winter months when heating demand is high.

- **Point emissions**: These are associated with elevated stacks, power plants, or other industrial facilities. In SIRANE, they are treated as vertical injections of pollutants above roof level and modeled using parameterizations or external plume dispersion models. The file `Sources_Ponctuelles.dat` includes metadata such as coordinates, height, temperature, and references to hourly-resolved emission profiles stored externally.
- **Background concentrations**: Rather than representing emissions per se, background values correspond to the inflow pollutant concentrations at the simulation domain boundaries. These are prescribed based on regional background levels or results from larger-scale chemical transport models. In the Lyon case study (Soulhac et al., 2012 [11]), background concentrations were set using measurements from suburban stations and remained constant throughout the simulation period, providing a baseline onto which local emissions were superimposed.

Figure 2.6: Location of point emission sources (black dots) in Turin.

Emission sources are temporally modulated using predefined hourly profiles. These modulation coefficients reflect typical daily emission cycles and differ between source types (e.g., traffic, heating, industrial activity). Each emission value defined in the input files is scaled by a time-varying coefficient for every hour of the year.
For instance, line and surface emissions in SIRANE are associated with hourly modulation factors such as:

- `Mod_Lin_0_NOX`: hourly scaling factor for NO_x;
- `Mod_Surf_NOX`: hourly scaling factor for surface NO_x emissions over the full domain.

There are seven different modulation coefficients that change based on the characteristics of the street. In the San Salvario district, all streets refer only to `Mod_Lin_0` coefficients. These coefficients are stored in dedicated modulation tables with 8760 rows (one per hour of the year). The values are derived from empirical or regulatory datasets, such as traffic counts or heating demand patterns, and are specific to each pollutant and emission source type.
For example, on January 1st, 2014:

- The average modulation coefficient for line NO_x was approximately **0.87**, peaking at **1.61** at 19:00;
- The modulation coefficient for surface NO_x remained **1.0** throughout the day.

These modulation values directly influence the final emission rates used in dispersion simulations and are fundamental to reproducing realistic diurnal concentration trends.

2.2.3 Output

The main output directories generated by SIRANE are `GRILLE` and `GRILLE_STATS`, which provide two-dimensional gridded estimates of pollutant concentrations across the simulation domain. The `GRILLE` folder contains hourly concentration fields computed for each pollutant over the entire simulation period. In contrast, the `GRILLE_STATS` directory stores statistical summaries—such as average, minimum, and maximum concentrations—aggregated over the full time window of the simulation.
Other relevant output folders include `RUES_PAR_HEURE` and `RUES_PAR_RUE`, both of which focus on street-level concentration outputs. The `RUES_PAR_HEURE` folder contains a series of `.dat` files, one per hour of simulation, where each row corresponds to a specific street segment (indexed from 0 to $N-1$) and reports the concentration at that time step.
The `RUES_PAR_RUE` folder, on the other hand, is structured as a set of files where each street index corresponds to a time series of pollutant concentrations. However, during the course of this work, it was observed that the values reported in `RUES_PAR_RUE` remain constant across all time steps, unlike the values in `RUES_PAR_HEURE`, which

show the expected temporal variation. This inconsistency suggests the presence of a bug in the source code affecting the `RUES_PAR_RUE` output.
Due to this issue and the lack of a direct built-in graphical representation of street-level concentrations on the urban network, this project focused primarily on the analysis of concentrations extracted from the `RUES_PAR_HEURE` files.

Figure 2.7: Example of a GRILLE output that represents NO_x in Turin.

2.2.4 Validation

In the Lyon validation case (Soulhac et al., 2017 [7]), the SIRANE model was applied over the entire urban agglomeration of Lyon for the year 2008. The simulation domain included over 21,000 street segments and accounted for temporally-resolved emission inventories, meteorological data, and detailed urban morphology. The validation process was carried out through comparison with both continuous automatic monitoring stations and passive diffusion tubes (PDTs).
Specifically, the model output was compared against:

- Hourly NO_2 concentration data from 15 automatic monitoring stations operated by the local air quality network (Atmo Auvergne-Rhône-Alpes). These stations were categorized as traffic, suburban, urban background, and industrial, allowing for performance assessment across a range of urban environments;
- Three passive sampling campaigns conducted over distinct periods of the year (February, May, October), involving over 70 sites distributed across the city. These PDTs provided spatially dense but temporally averaged concentration data.

To evaluate the model's performance, the following statistical indicators were used:

- **Fractional Bias (FB)** — a symmetric metric evaluating mean over- or under-prediction;
- **Normalized Mean Square Error (NMSE)** — assessing the variance between modeled and observed values;
- **Correlation coefficient (Corr)** — indicating the temporal co-variability of the signals;
- **Fraction of predictions within a factor of two (FAC2)** — representing the proportion of predictions that fall within a factor of two of observations.

Model performance was generally in agreement with the quality objectives established by Chang and Hanna (2004) [3], which recommend:

$$\begin{aligned} |\mathrm{FB}| &\leq 0.3 \\ \mathrm{NMSE} &\leq 4 \\ \mathrm{FAC2} &\geq 0.5 \\ \mathrm{Corr} &\geq 0.5 \end{aligned}$$

The SIRANE model reproduced the spatial distribution of NO_2 concentrations with good fidelity, particularly capturing elevated concentrations along major roadways and lower levels in peripheral or suburban areas. Temporal variability at traffic stations was also well represented, especially under neutral and unstable atmospheric conditions. Some discrepancies were observed under low wind or stable nighttime conditions, where vertical mixing is weak and local effects become dominant.
Notably, the use of passive diffusion samplers allowed for the identification of concentration gradients on a fine spatial scale. While these devices tend to slightly overestimate concentrations compared to reference analyzers, their high spatial density offered robust validation of the model's ability to resolve street-scale heterogeneity. Overall, the results confirmed that SIRANE can be effectively used for long-term, high-resolution simulation of NO_x and NO_2 concentrations over complex urban domains, making it a valuable tool for exposure assessment, policy design, and epidemiological studies.

Chapter 3

Urban Geometry

As briefly discussed in the previous chapter, the shapefile provides both the geographic (coordinates) and topological (arcs and nodes) framework within which SIRANE computes the diffusion and dispersion of pollutants. Urban geometry is therefore a fundamental component of the model and cannot be neglected when attempting to develop a predictive tool.
This dependency somewhat reduces the general-purpose applicability of the present work, since both the simulations and subsequent predictions are inherently "overfitted" to a specific shapefile — in this case, the district of San Salvario. As a result, the model must be retrained on a new dataset for each urban topology of interest, with dedicated SIRANE simulations tailored to the corresponding street network.
A truly general-purpose machine learning algorithm would need to learn the underlying geometric relationships between streets and be capable of adapting to arbitrary urban contexts once trained. However, the development of such an advanced model — which would likely require the integration of graph-based or geometric deep learning methods — lies beyond the scope of this thesis.

This chapter explains the rationale behind the dimensionality reduction approach adopted in this thesis, which allows focusing on the San Salvario district as a test case for surrogate model development.

3.1 Dimensionality reduction

By dimensionality reduction we refer to the spatial reduction of the original shapefile domain, limiting it to a smaller geographic region. The rationale behind this simplification is to significantly reduce the computational cost associated with the simulations, allowing for faster execution times. Furthermore, since this work adopts a highly experimental and data-driven methodology, the idea was to first validate

the approach on a smaller-scale "toy model" before extending it to the entire urban area. Working on a reduced domain enables rapid iteration, facilitates testing of different modeling techniques, and simplifies debugging. Given the limited and fixed timeframe of this research, this strategy represents a major advantage.

3.2 Methodology used

To crop the shapefile and isolate the desired urban area (San Salvario district), the open-source GIS software QGIS was employed. After importing the original street network shapefile, all streets and nodes lying outside the region of interest were manually deleted. This manual pruning alters the urban geometry while preserving the topological structure and attribute fields of the remaining segments.
However, this process affects the correspondence between street identifiers and emission fluxes. In SIRANE, line emissions are assigned sequentially to street segments according to their internal index in the shapefile. Therefore, once the domain is reduced, the linear emission file must be updated accordingly to maintain consistency. This requires:

- identifying the original indices corresponding to the remaining streets in the San Salvario district;
- removing emission records for deleted streets;
- reindexing the reduced emission file from 0 to $N - 1$, where N is the new number of retained streets.

An important consideration concerns the handling of emissions associated with streets that are removed during the domain reduction process. Simply assigning a value of zero to their emissions in the input file is not sufficient. This is because SIRANE links emission values to street segments based solely on their index order in the shapefile. If the emission file is not updated to reflect the new reduced geometry, the first N emission values (where N is the number of streets in the reduced shapefile) will be incorrectly assigned to the remaining segments, regardless of their actual identity in the original network. This can lead to erroneous spatial distributions, particularly if zero values (originally associated with now-removed streets) are wrongly attributed to valid segments in the reduced model. To avoid this misalignment, it is therefore essential to remove the emission entries corresponding to the deleted streets and to reindex the emission file in a way that aligns precisely with the geometry of the reduced shapefile. This ensures reproducibility and prevents mismatches between the geometric and emissions datasets during the simulation phase.

3.3 Comparative analysis

Once a suitable domain reduction technique has been identified, it becomes possible to perform a comparative simulation over both the full city network and a reduced domain focused on San Salvario. The primary goal is to evaluate how the truncation of the original shapefile affects the predicted concentrations within the district, and to assess the extent to which emissions from the rest of the urban area influence local pollutant levels. To this end, a dedicated simulation was first conducted over the complete urban domain. This preliminary step was designed to investigate the interactions between San Salvario and the surrounding city, providing insight into the spatial scale and magnitude of pollutant exchange across district boundaries.

3.3.1 Interaction between San Salvario and the rest of the city

The experimental setup consists of a one-day simulation over the full city domain, using January 1st, 2014 as a representative test case. After the simulation, the pollutant concentrations computed for the streets within the San Salvario district are extracted, and the spatial average and time average are calculated —i.e., the average NO_x concentration across all streets at each hour and the average NO_x across 24 hours at each street.
To better understand the relative influence of local versus non-local emissions, four emission scenarios for linear (traffic-related) sources were considered:

- **FLTR**: This is the reference case, in which linear emissions are present across the entire domain, both inside and outside the San Salvario district;
- **FLTR_SS**: In this configuration, linear emissions are active only within San Salvario. All street segments outside the district have their emissions set to zero;
- **FLTR_NOSS**: This scenario is complementary to the previous one: linear emissions are present only outside San Salvario, and they are set to zero within the district itself;
- **FLTR_0**: In this final scenario, all linear emissions are set to zero, both inside and outside San Salvario. The only contributions to street-level concentrations come from surface, point, and background sources.

In all four scenarios, the focus remains on the concentrations predicted within the San Salvario district, allowing an assessment of how emissions in different parts of the city affect the local air quality.
The images below (Fig. 3.1) display the gridded concentration output generated by SIRANE for the four emission scenarios previously described. Although concentration differences at the individual street level are not easily discernible from these plots, a clear trend emerges: linear (traffic-related) emissions play a dominant role in determining overall concentration levels within the domain. When these emissions are deactivated, the resulting concentrations are primarily attributable to the remaining sources—namely surface, point, and background emissions.
This observation highlights an important aspect of pollutant dynamics within the urban canopy. The transport and diffusion of pollutants along the street axis appear to have a predominantly local effect, influencing adjacent or directly connected street segments. However, their impact diminishes rapidly with distance, suggesting that inter-street pollutant exchange does not significantly affect remote parts of the domain. This supports the hypothesis that the influence of street-level emissions is spatially limited and largely confined to the immediate surroundings of the emission source.

(a) Scenario FLTR

(b) Scenario FLTR_SS

(c) Scenario FLTR_NOSS

(d) Scenario FLTR_0

Figure 3.1: Grille output of SIRANE for the averadge NO_x concentration on January 1st, 2014 in the four different scenarios considered: FLTR (linear emissions across all the streets), FLTR_SS (linear emissions only in San Salvario's street), FLTR_NOSS (linear emissions only outside San Salvario), FLTR_0 (no linear emissions).

We now present a set of numerical results obtained using MATLAB that further explore the concentration dynamics within the San Salvario district. Fig. 3.2 shows the progressive reduction in the spatially averaged NO_x concentration over the district as linear emissions are progressively removed—first from the streets outside San Salvario, and then from within. Although concentration differences at the individual street level are not easily distinguishable in gridded plots, the effect of linear emissions on overall concentration levels is clearly evident.
The highest concentrations are observed in the scenario where all streets—both within and outside the San Salvario district—contribute with active emissions (**FLTR**). Concentration levels decrease when emissions are restricted to streets located exclusively within San Salvario (**FLTR_SS**), and are further reduced in the scenario where all street-level emissions are completely suppressed (**FLTR_0**). In the case where only streets outside San Salvario emit pollutants (**FLTR_NOSS**), the resulting concentrations within the district are higher than those observed in **FLTR_0**, highlighting the influence of pollutant advection from the surrounding urban area. This behavior is physically consistent and confirms the central role of local traffic emissions in determining street-level concentration patterns, while also indicating that inter-district transport has a secondary but non-negligible effect.

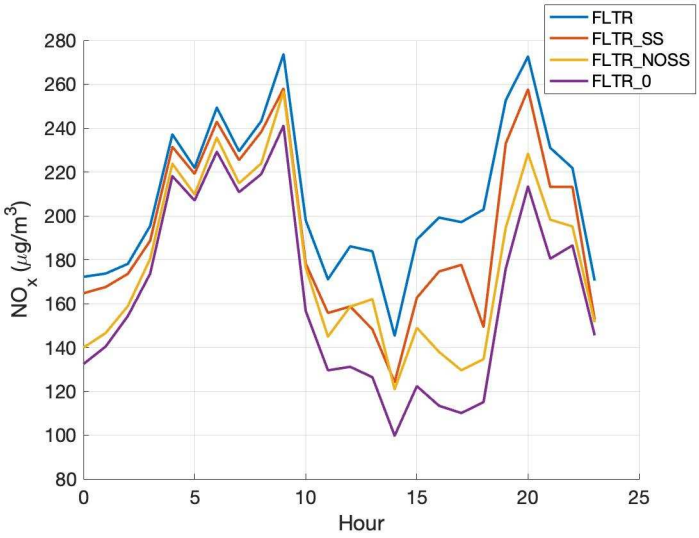


Figure 3.2: Spatially averaged NO_x concentration time series over San Salvario on January 1st, 2014 for every hour of simulation in the four different scenarios considered: FLTR (linear emissions across all the streets), FLTR_SS (linear emissions only in San Salvario's street), FLTR_NOSS (linear emissions only outside San Salvario), FLTR_0 (no linear emissions).

The images below (Fig. 3.3) presents a complementary view: the time-averaged NO_x concentration on January 1st, 2014, for each street in the district. The same decreasing trend is observed across the four emission scenarios.

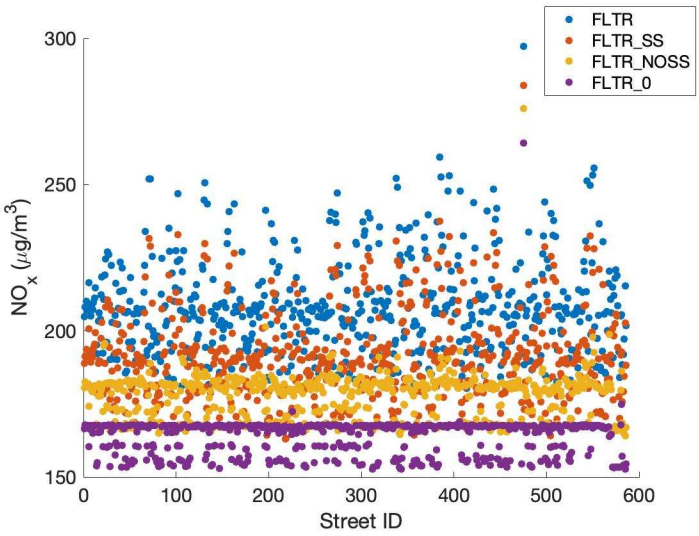
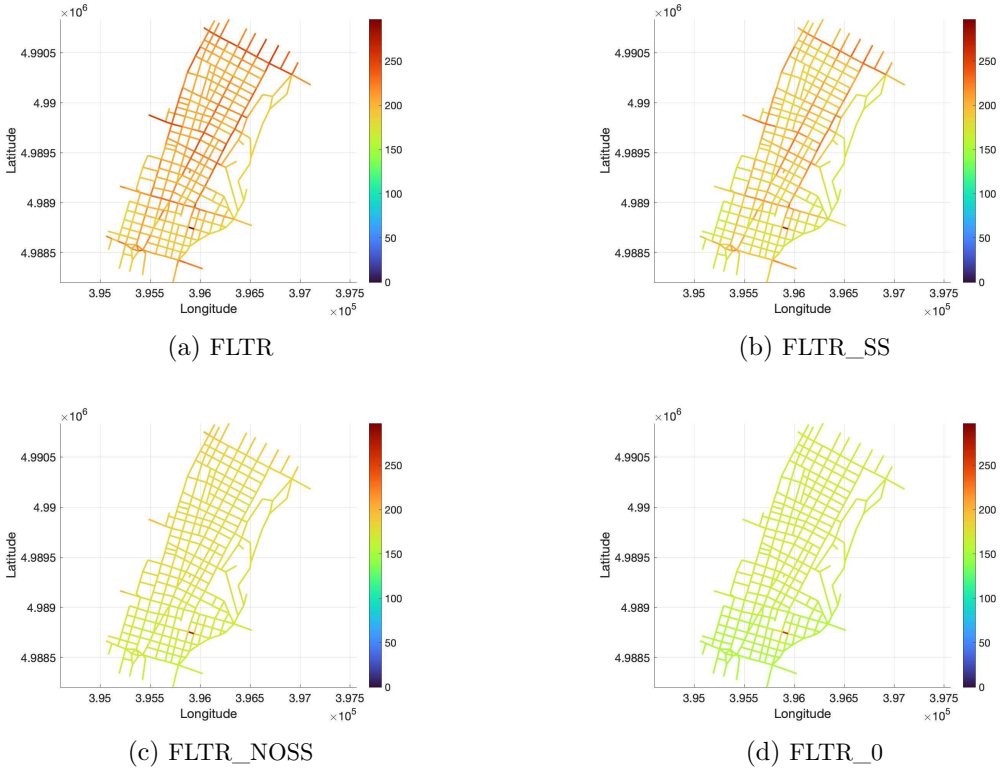


Figure 3.3: Time-averaged NO_x concentration on January 1st, 2014 for each street in San Salvario in the four different scenarios considered: FLTR (linear emissions across all the streets), FLTR_SS (linear emissions only in San Salvario's street), FLTR_NOSS (linear emissions only outside San Salvario), FLTR_0 (no linear emissions).

To further clarify these trends, the time-averaged concentrations are mapped directly onto the San Salvario street network in Fig. 3.4.

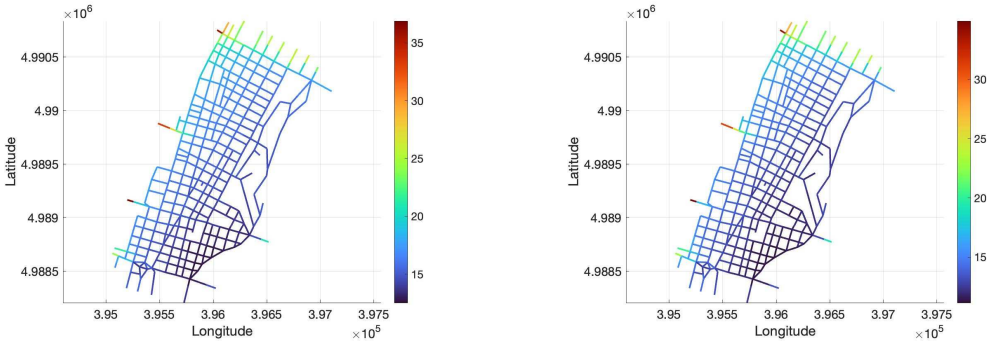


(a) FLTR (b) FLTR_SS

(c) FLTR_NOSS (d) FLTR_0

Figure 3.4: Time-averaged NO_x concentration on January 1st, 2014 for each street in San Salvario in the four different scenarios considered: FLTR (linear emissions across all the streets), FLTR_SS (linear emissions only in San Salvario's street), FLTR_NOSS (linear emissions only outside San Salvario), FLTR_0 (no linear emissions).

Finally, the plots in Fig. 3.5 provide a numerical representation of the local effect of pollutant transport, consistent with previous findings from the SIRANE gridded output. They show the difference in time-averaged concentrations between two pairs of complementary scenarios: FLTR vs. FLTR_SS and FLTR_NOSS vs. FLTR_0.



(a) FLTR − FLTR_SS

(b) FLTR_NOSS − FLTR_0

Figure 3.5: Differences in time-averaged NO_x concentrations between complementary emission scenarios on January 1st, 2014 for each street in San Salvario. Picture (a) shows the difference in concentretion level between scenario FLTR (linear emissions across all the streets) and FLTR_SS (linear emissions only in San Salvario's street). Picture (b) shows the difference in concentration level FLTR_NOSS (linear emissions only outside San Salvario) and FLTR_0 (no linear emissions).

As shown in Fig. 3.5, the most significant concentration differences occur along the peripheral segments that connect the district to the surrounding urban fabric. This indicates that pollutant transport has a predominantly local effect, strongly influencing adjacent streets but quickly decaying with distance.

3.3.2 Comparison between full and truncated domain

This section focuses on a numerical comparison between simulations carried out on the full domain and those restricted to the truncated domain. Based on the results presented in the previous section, one would expect lower concentration levels in the truncated domain due to the smaller number of active streets and the absence of emissions originating from outside San Salvario.

As shown in Fig. 3.6, this expected trend is clearly confirmed. Moreover, the simulation over the truncated domain reveals a different temporal pattern in the spatially averaged concentration, indicating that emissions from the rest of the city influence not only the absolute concentration levels but also their diurnal modulation. The truncated domain simulation appears more regular, with reduced variability and fewer sharp concentration peaks, reinforcing the role of the broader urban network in amplifying transient episodes.

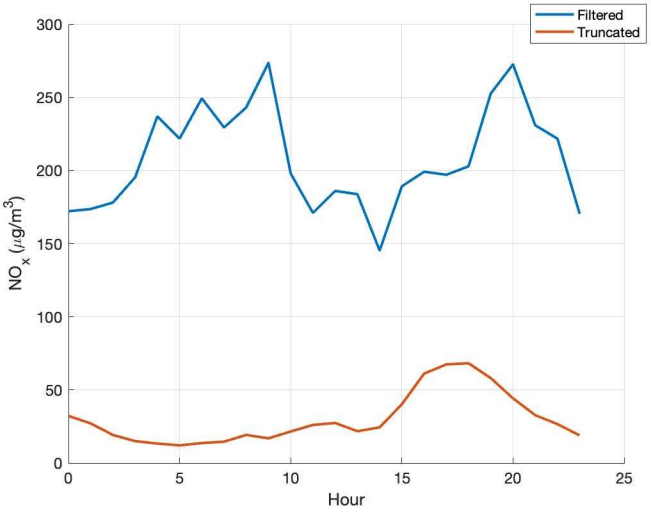


Figure 3.6: Spatially averaged NO_x time series concentration comparison between the full and truncated domain simulations over San Salvario on January 1st, 2014 for every hour of simulation. The blue line represents the average concentration in San Salvario with a simulation on the full domain. The red line represents the average concentration in San Salvario with a simulation on the truncated domain.

At the street level, a comparison of the time-averaged NO_x concentrations (Fig. 3.7) reveals a similar spatial pattern in both domains, albeit with lower absolute values in the truncated case. Once again, reduced variability is observed in the truncated domain, suggesting that emissions from outside the district contribute significantly to the dynamic range of concentrations within San Salvario.

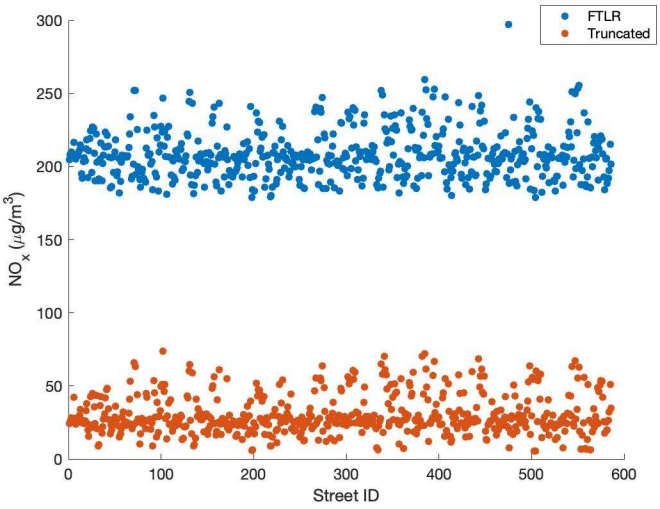
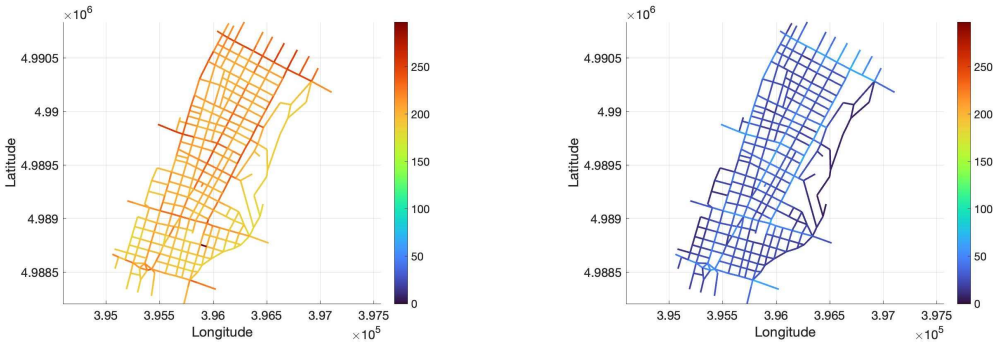


Figure 3.7: Time-averaged NO_x concentration comparison between the full and truncated domain simulations over San Salvario on January 1st, 2014 for each street in San Salvario. The blue scatter plot represents the average concentration in one day with a simulation on the full domain. The red scatter plot represents the average concentration in one day with a simulation on the truncated domain

A spatial visualization of the time-averaged concentrations (Fig. 3.8) further supports this observation: although concentration levels are generally lower in the truncated domain, the spatial distribution of pollutant hotspots remains consistent. Streets with higher concentrations in the full domain simulation also exhibit higher values in the truncated simulation, albeit scaled down. This confirms that the truncated domain can still capture the relative spatial patterns of pollution across the district.



(a) FLTR (Full Domain)

(b) TRCT (Truncated Domain)

Figure 3.8: Time-averaged NO_x concentration across San Salvario streets in the full and truncated domain simulations on January 1st, 2014 for each street in San Salvario. Panel (a) shows the concentretion level for a simulation on the full domain. Panel (b) the concentretion level for a simulation on the truncated domain.

The similarity in spatial distribution between the two domains lends credibility to the use of the truncated domain for further analysis. The key advantage of this approach, however, lies in the significant computational savings, as summarized in Table 3.1.

Execution Component	**Full Domain (s)**	**Truncated Domain (s)**	**Speed-up (%)**
Global execution time	2948.071	702.560	76.2%
Iteration time	2839.743	677.865	76.1%
Avg. time per iteration	118.323	28.244	76.1%
Initialisations	119.730	109.455	8.6%
Puffs	57.154	19.126	66.5%
Point sources	44.902	19.924	55.6%
Retrotrajectories	324.791	252.257	22.3%
Street plume	2037.620	71.256	96.5%
Chemistry	13.240	11.239	15.1%
Grid summation	50.432	48.143	4.5%
File writing	186.050	142.784	23.3%
Releases	5.824	3.681	36.8%

Table 3.1: Comparison of execution times between full and truncated domain simulations.

The computational advantage is substantial, especially given the limited and fixed timeframe of this research. For this reason, all subsequent analyses—namely the sensitivity study and the training of the predictive model—will be carried out using simulations performed on the truncated domain.

3.4 Test dataset

The test dataset adopted throughout this work corresponds to a year-long simulation over the reduced domain of San Salvario, using real and validated meteorological inputs along with consistent emissions and temporal modulation coefficients. This dataset was originally generated and employed by Matteo Bo in his PhD thesis *Study of aerosols air pollution assessments in indoor and outdoor environments based on measuring and modelling approaches* [1]. Its relevance lies in the fact that, in the absence of real-life measurements of pollutant concentrations at street level, the

outputs provided by SIRANE during this specific simulation period (i.e., the year 2014) will serve as the reference benchmark against which the performance of all subsequent predictive models will be validated.

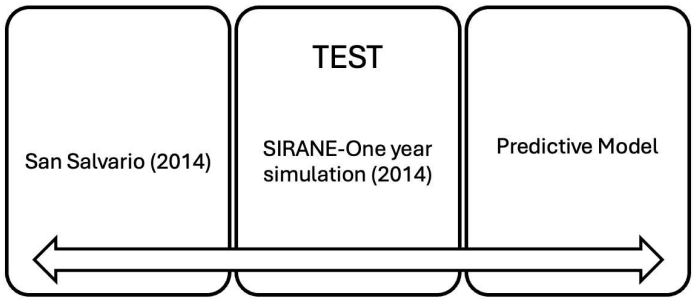


Figure 3.9: Logic relationship between the three different universes. The real world is visible to predictive model only throught the SIRANE output of one year simulation on San Salvario.

In addition to its role in model validation, the test dataset also proved fundamental in the exploratory analysis of the meteorological parameters. Specifically, it enabled the identification of the realistic range of variability (upper and lower bounds) for key inputs such as wind direction, air temperature, friction velocity (u_*), and inverse Monin-Obukhov length (L_{MO}^{-1}). This analysis guided the definition of the phase space resolution used to construct the synthetic database presented in Chapter 4, ensuring that the artificial input combinations were both representative and physically plausible.

Meteorological parameter distribution

The test dataset comprises 8760 data points, corresponding to each hour of simulation over the year 2014. Among these, only 8758 correspond to unique meteorological parameter combinations; two specific combinations appear twice, as detailed in Table 3.2.

Hour of Simulation	φ	T	u_*	L_{MO}^{-1}
Hour 461 (20/01/2014 04:00)	236.5	5.2	0.08	0.0111
Hour 934 (08/02/2014 21:00)	236.5	5.2	0.08	0.0111
Hour 765 (01/02/2014 20:00)	259.3	5.3	0.08	0.0111
Hour 1080 (14/02/2014 23:00)	259.3	5.3	0.08	0.0111

Table 3.2: Duplicated meteorological parameter combinations in the test dataset.

Fig. 3.10 below shows the empirical distribution of the four meteorological variables used in the simulations. This analysis provides key insights into the range, density, and variability of the input space explored through the year-long simulation.

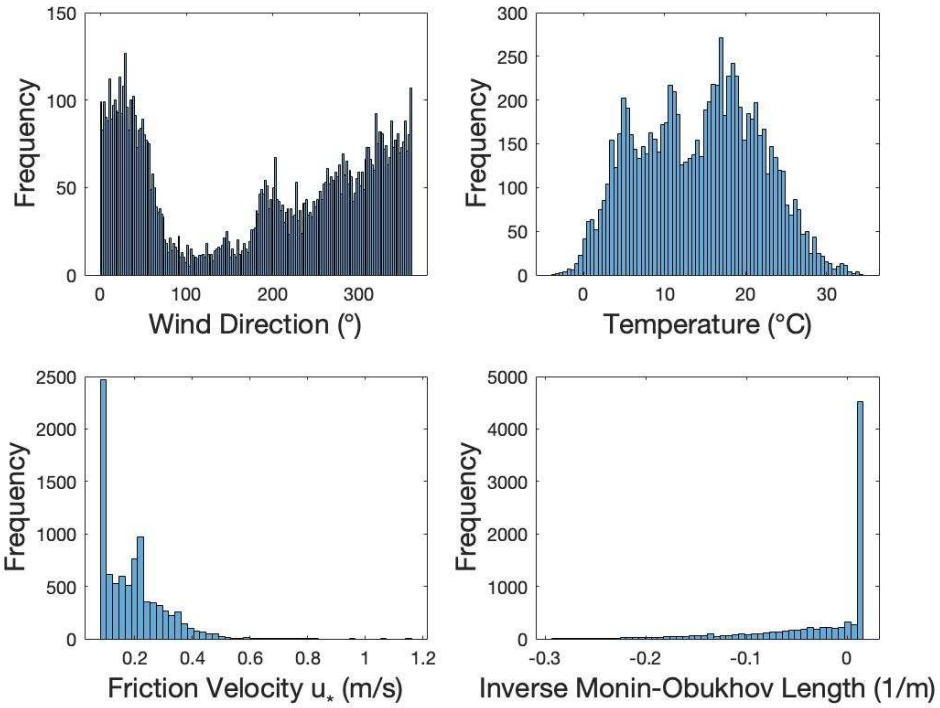


Figure 3.10: Empirical distribution of the four meteorological parameters in the test dataset.

From the figure, it is evident that both the friction velocity (u_*) and the inverse Monin-Obukhov length (L_{MO}^{-1}) exhibit highly skewed distributions. The former is heavily concentrated around low values, while the latter shows a strong peak around the value 0.0111, often associated with stable stratification conditions. Conversely, temperature (T) and wind direction (φ) show broader and more uniform distributions.

These findings highlight a potential strong statistical dependence between u_* and L_{MO}^{-1}, a relationship that will be further explored in Chapter 4. Moreover, understanding these distributions was crucial in selecting appropriate bounds and discretization steps for the synthetic phase-space construction described in the following chapters.

Modulation coefficients time series

As explained in Chapter 2, both street and surface emissions are subject to temporal modulation through specific coefficients that capture the variation of emission intensity over different hours of the day, days of the week, and months of the year. In the case of surface emissions, the modulation coefficient is constant and equal to 1, reflecting a steady emission profile. Conversely, as shown in Fig. 3.11, the

modulation coefficient for street emissions exhibits significant temporal variability, influenced by traffic patterns and diurnal cycles.

This temporal modulation presents a challenge in the construction of the synthetic dataset used for training the predictive model. Since the artificial dataset is designed to be meteorology-dependent but not time-dependent, the temporal component of the emission modulation must be carefully managed. While this issue does not affect the sensitivity analysis phase—where all inputs are treated equally—the lack of time-representative modulation can result in a mismatch between model predictions and real-world scenarios when evaluating model performance against the test dataset.

To address this discrepancy and ensure consistency between the synthetic dataset and the real-case reference, a dedicated strategy was implemented during model development. This aspect will be discussed in greater detail in the chapter 5, where the integration of temporal modulation into the learning process will be presented.

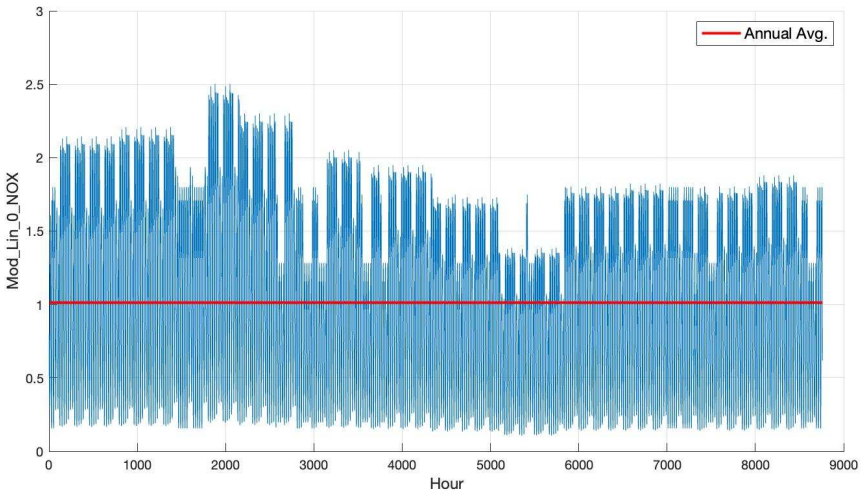


Figure 3.11: Time series of the modulation coefficient for NO_x street emissions in San Salvario. The blue lines is the annual variation of the coefficient for each simulation's hour. The red horizontal line is the annual average.

Chapter 4

Sensitivity Analysis

Once the modeling domain is defined, an essential step is to understand how sensitive pollutant concentrations are to varying meteorological conditions. This sensitivity analysis provides the foundation for developing accurate predictive models.

This chapter presents the methodology and results of a sensitivity analysis aimed at characterizing how key meteorological parameters influence street-level pollutant concentrations. The insights gained here are instrumental for guiding the design of the surrogate models discussed in the following chapter.

4.1 Outline of the work

The first step in the modelling framework consists in creating a database that stores the street-level NO_x concentrations for the San Salvario district under different meteorological scenarios. The choice of focusing primarily on NO_x stems from its relative chemical stability: in contrast to ozone or other reactive species, NO_x (defined as the sum of NO and NO_2) is not subject to significant transformation processes over short timescales. This characteristic allows us to treat it as a passive scalar, thus avoiding the introduction of non-linear reaction dynamics which would complicate the learning task for statistical models. SIRANE includes a simplified chemical module to account for nitrogen oxide reactions involved in ozone formation and destruction. This chemical mechanism is based on the well-known Chapman cycle:

$$\begin{cases} NO_2 + h\nu \xrightarrow{k_1} NO + O\cdot \\ O\cdot + O_2 \xrightarrow{k_2} O_3 \\ NO + O_3 \xrightarrow{k_3} NO_2 + O_2 \end{cases} \tag{4.1}$$

where k_1, k_2, and k_3 are the kinetic constants associated with each reaction.

Despite the inclusion of this chemical module, the modelling approach used in this work relies on a simplification. First, SIRANE simulates the dispersion of NO_x as if it were a passive tracer. Subsequently, assuming photo-stationary equilibrium, it derives the concentrations of NO, NO_2, and O_3 from the NO_x values through the above chemical scheme. This approach justifies our choice to model NO_x directly, as it simplifies the physics while still retaining an interpretable and relevant quantity for street-level air quality analysis.
Once the database has been set up, the next step involves analysing how the mean concentration levels within the district vary as a function of key meteorological variables. In particular, we focus on the influence of wind direction (φ), air temperature (T), friction velocity (u_*), and the inverse of the Monin–Obukhov length (L_{MO}^{-1}). The analysis is restricted to the spatial average of the NO_x concentrations across the San Salvario district, providing a global overview of how each meteorological condition modulates pollutant levels in the area.

4.2 The database creation

As discussed in Chapter 2, SIRANE considers four types of pollutant emission sources: linear, surface, point, and background. In the construction of our database, we chose to include only the first two—linear and surface emissions—while neglecting point and background sources. This decision was motivated by two main factors: first, preliminary analyses on the test dataset showed that point sources contributed negligibly to time-based statistics (such as hourly or daily averaged concentrations); second, excluding these components allowed for a substantial reduction in computational cost, which was essential given the limited timeframe of the project. Regarding background concentrations, these were deliberately excluded as they represent a scalar quantity uniformly added across the simulation domain and are not subject to dispersion processes within the canopy model. As such, they act as a constant offset on top of the street-level concentrations and do not carry information about spatial variability or local emission characteristics. Including background concentrations in the synthetic dataset would have introduced a constant bias across all data points, offering little value to the learning process of a data-driven model that aims to understand and reproduce the effects of local dynamics. However, subsequent analysis revealed that the assumption of neglecting point sources holds primarily for temporal statistics (Fig. 4.1), and not necessarily for spatial metrics. While point sources may have minimal impact on time-averaged values, they can produce localized peaks that influence the spatial distribution of concentrations. Unfortunately, by the time this limitation was identified, the database had already been generated

and the simulation results finalized, leaving no opportunity for further adjustment within the available schedule.
A more detailed discussion of this limitation, along with its implications and possible corrections, will be presented in the following chapter.

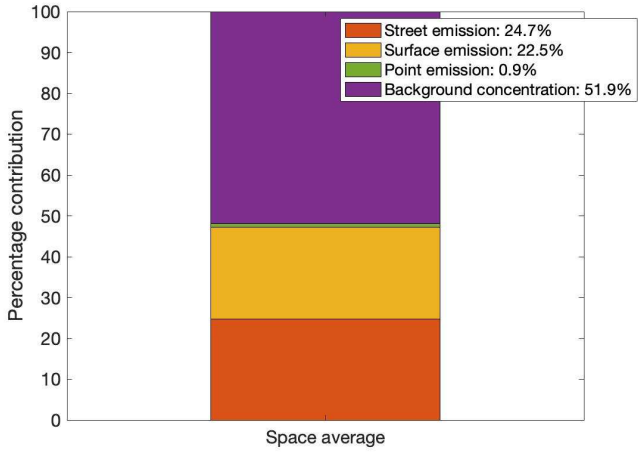


Figure 4.1: Barplot that shows the percentage contribution on the spatially avaraged annual concentration of the different source emission. Red part represents the street emission contribution, yellow one the surface, green one the point and purple is the background concentration contribution.

4.2.1 Phase space for meteorological inputs

Based on the empirical analysis of the meteorological parameter distribution in the test dataset, a structured discretization of the input space was defined to build the synthetic phase space used in Chapter 4. The discretization criteria are as follows:

- **Wind direction** φ: uniformly sampled from 0° to 360° with a step size of 5°, yielding a total of 73 distinct values. This discrete set is denoted by Θ.
- **Air temperature** T: sampled in the range from −10°C to 45°C with increments of 5°C, resulting in 12 values. This discrete set is denoted by $\mathcal{T}$.
- **Friction velocity** u_*: discretized from 0.05 to 1.2 m/s using a step of 0.05 m/s, producing 24 unique values. This discrete set is denoted by Φ.

The inverse Monin-Obukhov length L_{MO}^{-1}, being a derived stability parameter that does not vary naturally on a regular linear scale, was discretized using a manually, as in (**Soulhac et al., 2012**[11]), selected set of representative nine values that span a wide range of atmospheric stability regimes; the following set is denoted by $\mathcal{I}$:

- **Inverse Monin-Obukhov length L_{MO}^{-1}**:

$$-0.3,\ -0.2,\ -0.1,\ -0.004,\ -0.019,\ 0,\ 0.004,\ 0.019,\ 0.05,\ 0.12$$

The total number of unique meteorological parameter combinations, based on the discretized phase space described above, is:

$$N_{\text{comb}} = 73 \times 12 \times 24 \times 9 = 189216$$

This means that in order to construct the complete synthetic dataset, a total of 189216 hourly simulations ($\sim$ 21.6 years) are required, each corresponding to a distinct combination of wind direction, temperature, friction velocity, and inverse Monin-Obukhov length. This exhaustive sampling of the input parameter space enables the construction of a high-resolution dataset. In this initial configuration, all meteorological parameters are treated as mutually independent, and every possible combination within the defined ranges is considered feasible. This assumption enables the exhaustive exploration of the phase space but does not necessarily reflect realistic atmospheric conditions. In Section 4.2.3, we will introduce a refinement of this approach, where only physically plausible meteorological scenarios—based on observed correlations and joint distributions—are retained in order to improve the representativeness of the synthetic dataset.

4.2.2 Data structure

Once all simulations were completed, the concentration levels for each street were stored in a five-dimensional array denoted as C, with dimensions:

$$D_{\text{C}} = 586 \times 73 \times 12 \times 24 \times 9$$

where the first dimension indexes the 586 streets in the San Salvario district, and the remaining four dimensions correspond respectively to the discretized values of wind direction, air temperature, friction velocity, and inverse Monin-Obukhov length.
Based on this core array, additional four-dimensional arrays were constructed by applying spatial statistics across the street index (first dimension), yielding the following derived datasets:

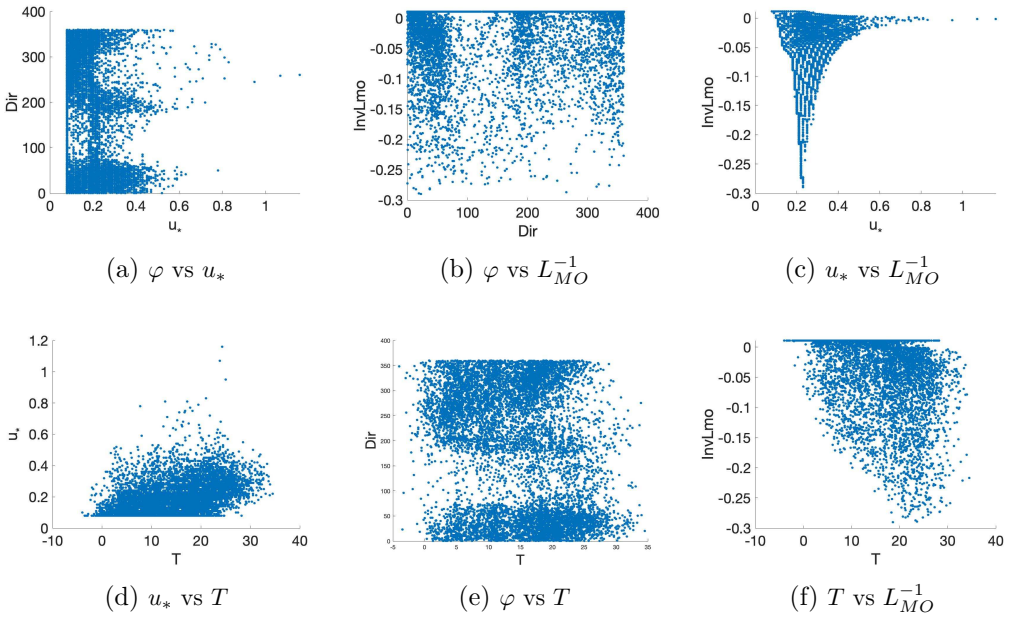
- C_{mean}: spatial average of the concentration for each combination of meteorological parameters;
- C_{min}: minimum concentration observed across all streets for each parameter set;

- $C_{\max}$: maximum concentration observed across all streets for each parameter set;
- C_{var}: variance of concentration levels across all streets for each parameter set.

The sensitivity analysis was conducted on all of these derived datasets; however, in this thesis we report only the results concerning C_{mean}, since the primary objective of the analysis was to investigate how meteorological variables influence average concentration levels across the district. Pointwise statistics such as minimum and maximum concentrations, being more susceptible to localized or episodic effects, were deemed less informative in this context.

4.2.3 Feasible weather

This section investigates the interdependencies among meteorological parameters and their implications for sensitivity analysis. While the synthetic dataset was initially constructed by assuming complete independence between variables, it is well known that certain combinations of meteorological conditions are physically unrealistic and should therefore be excluded from meaningful analysis.



(a) φ vs u_* (b) φ vs L_{MO}^{-1} (c) u_* vs L_{MO}^{-1}

(d) u_* vs T (e) φ vs T (f) T vs L_{MO}^{-1}

Figure 4.2: Empirical pairwise distributions of the four meteorological parameters. Each point in these scatter plots is a pair of meteorological parameters sampled from the output of SIRANE's one year simulation over San Salvario.

Fig. 4.2 suggests that most meteorological variables appear relatively uncorrelated, with the notable exception of the relationship between friction velocity u_* and the

inverse Monin-Obukhov length L_{MO}^{-1}. This correlation stems from the theoretical relationship:

$$L_{MO} = -\frac{\rho c_p \theta u_*^3}{\kappa g H_0}$$

According to this expression, large values of u_* are physically plausible only under neutral or near-neutral stability conditions, while under stable or unstable stratification, lower values of u_* are more common. This phenomenon is clearly illustrated in Fig. 4.1.

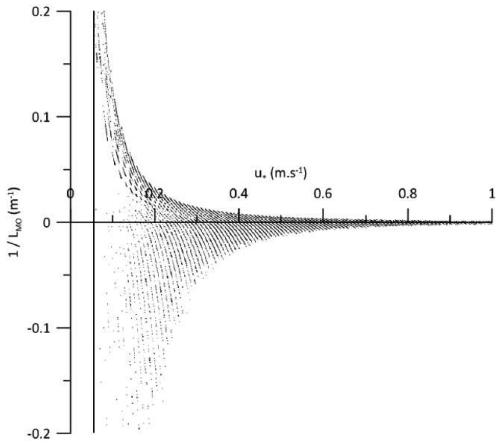


Figure 4.3: Distribution of meteorological data in the (L_{MO}^{-1}, u_*) space, independent of wind direction.

To account for this physical constraint, the original five-dimensional dataset constructed under the independence assumption was filtered to retain only meteorologically feasible combinations of (L_{MO}^{-1}, u_*) and the relative. The accepted region is delineated in red in Fig. 4.3.

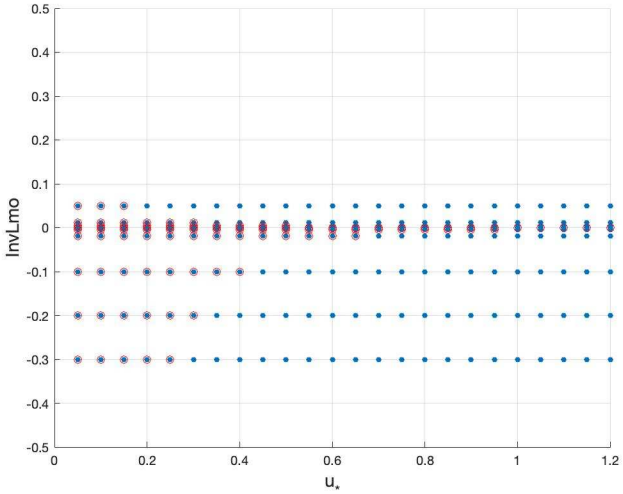


Figure 4.4: Feasibility region of the dataset: red dots represent meteorologically plausible (L_{MO}^{-1}, u_*) pairs. The dataset is filtered considering only points in the feasibility region.

In order to maintain the same data structure for consistency in subsequent analyses, all infeasible parameter combinations were assigned `NaN` values. This strategy ensured that both the sensitivity analysis and model training stages were conducted using only valid meteorological scenarios, minimizing the influence of outliers or physically unrealistic data points.

4.3 Analysis

In this section, we investigate how key meteorological parameters—namely temperature, friction velocity, and wind direction—affect pollutant concentrations at street level under varying atmospheric stability regimes. Each parameter is analysed independently to understand its specific contribution to the spatially averaged NO_x concentration across the district. The objective is to quantify the sensitivity of the system to these variables and identify the most influential ones, providing insight into the physical mechanisms driving pollutant dispersion in the urban environment.

4.3.1 Temperature

The first meteorological variable considered in the sensitivity analysis is the air temperature T. In the context of this study, NO_x is treated as a passive scalar, meaning it does not undergo chemical transformations that are directly influenced by temperature. Nevertheless, temperature can affect dispersion dynamics indirectly, primarily through its influence on atmospheric turbulence and boundary layer processes.

The atmospheric boundary layer height h—defined as the lowest part of the troposphere that is directly affected by surface forcings—can vary with temperature, particularly under unstable conditions. Such variations may influence turbulent transport and vertical mixing through the implicit dependency of the vertical turbulent velocity component σ_w on the boundary layer height h, as shown in Eq. 2.5 and the related formulations in Eq. 4.2. This, in turn, could affect pollutant concentrations at street level.

$$\begin{cases} \sigma_w = \sqrt{\sigma_{wc}^2 + \sigma_{wn}^2} & \text{if } L_{MO} < 0 \text{ - unstable} \\ \text{with } \sigma_{wc} = \sqrt{0.4}\, w_* \, 2.1(z/h)^{1/3}(1 - 0.8z/h) & \text{and } \sigma_{wc} = 1.3u_*(1 - 0.8z/h) \\ \sigma_w = 1.3u_*(1 - 0.8z/h) & \text{if } L_{MO} \to \infty \text{ - neutral} \\ \sigma_w = 1.3u_*(1 - 0.5z/h)^{3/4} & \text{if } L_{MO} > 0 \text{ - stable} \end{cases} \tag{4.2}$$

However, the actual extent to which temperature modulates NO_x concentrations at street level remains uncertain and is precisely one of the aspects investigated in this sensitivity analysis.

The plot in Fig. 4.5 is obtained by averaging the spatially averaged concentration values stored in C_{mean} over the sets Θ (wind directions) and Φ (friction velocities), for each fixed value of temperature and inverse Monin-Obukhov length $i \in \mathcal{I}$:

$$\overline{C_{\text{mean}}(T)} = \frac{1}{|\Theta||\Phi|} \sum_{Dir \in \Theta, u_* \in \Phi} C_{\text{mean}}(Dir, T, u_*, i) \qquad \forall i \in \mathcal{I} \tag{4.3}$$

Two main observations emerge from the analysis:

- The concentration levels remain nearly constant as temperature varies, with only a slight decrease observable under the most unstable conditions.
- An oscillatory pattern is evident across different stability regimes: concentrations tend to decrease from unstable to neutral conditions, and then increase again moving towards more stable scenarios.

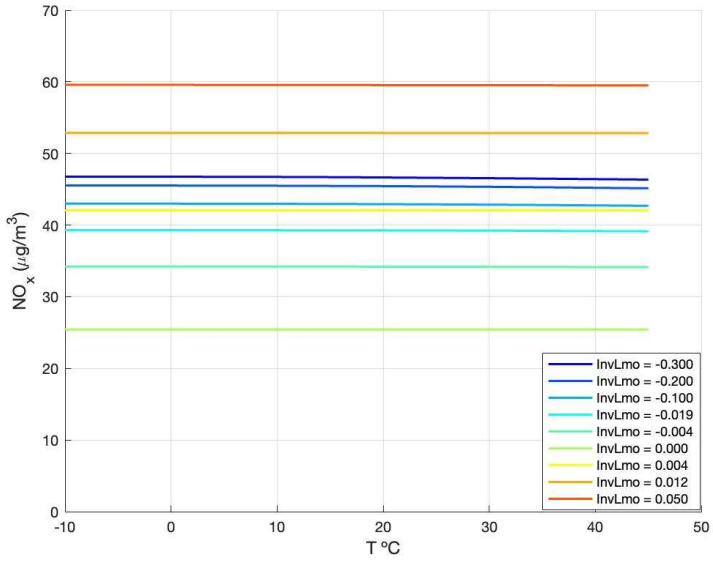


Figure 4.5: Variation of spatially averaged NO_x concentration levels as a function of temperature for all nine different stability regimes. Lines transition in color from dark blue to light green as conditions go from most unstable to neutral, and from light green to red as conditions shift from neutral to most stable.

Figure 4.6 further illustrates the relative variation in concentration between the minimum and maximum temperature values for each stability class. The largest variation is observed under highly unstable conditions, yet it remains below 1%.

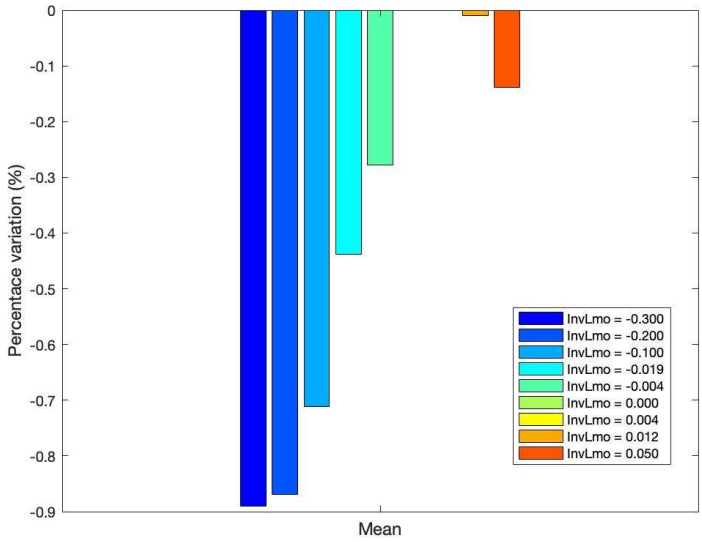


Figure 4.6: Percentage variation in NO_x concentration levels across the temperature range for all nine different stability regimes. Bars from dark blue to light green go from most unstable to neutral. Bars from light green to red go from neutral to most stable.

While the oscillatory behaviour described in the second point may depend on the interactions with other meteorological parameters (analysed in the following sections), the overall impact of temperature on spatially averaged concentration levels appears negligible. As a result, temperature can be excluded from the dataset without significant loss of information, reducing the dimensionality of the database by a factor of 12 and simplifying subsequent analyses.

4.3.2 Friction velocity

We expect friction velocity to be the most influential variable in regulating street-level concentrations, as it directly affects both the advective transport along the street axis through the term U_H in Eq. 2.3 and the turbulent vertical exchange via the parameter σ_w, as shown in Eq. 4.2. Since the temperature dimension has been removed from the dataset, the average concentration profile in Fig. 4.7 is computed by averaging over the set of wind directions Θ:

$$\overline{C_{\text{mean}}(u_*)} = \frac{1}{|\Theta|} \sum_{Dir \in \Theta} C_{\text{mean}}(Dir, u_*, i) \qquad \forall i \in \mathcal{I} \tag{4.4}$$

Several key observations can be drawn from Fig. 4.7:

- As friction velocity increases, the average concentration levels decrease significantly across all stability regimes.

- Due to the constraints imposed by meteorological feasibility (see Fig. 4.3), high values of u_* occur only under neutral or near-neutral conditions.
- An unexpected behavior is observed in the lower range of friction velocity values, approximately within $[0.05, 0.2]$ m/s, where unstable conditions correspond to higher concentration levels compared to more stable scenarios—contrary to what physical intuition would suggest.

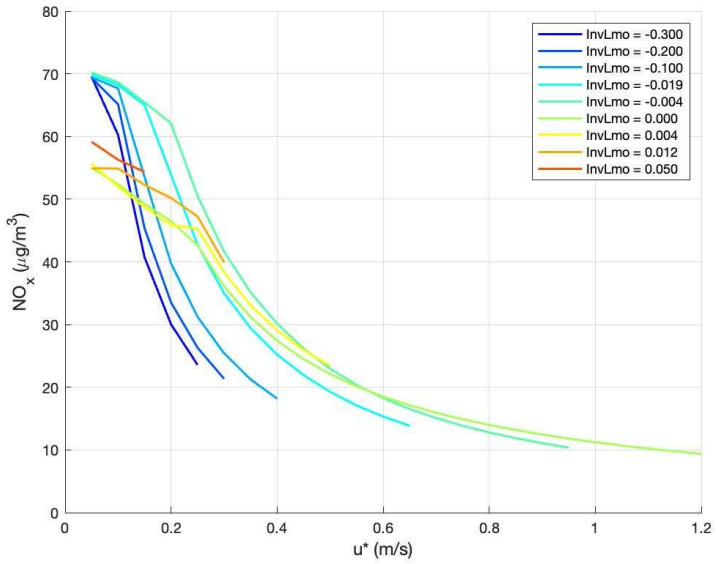


Figure 4.7: Variation of spatially averaged NO_x concentrations as a function of friction velocity for all nine different stability regimes. Lines transition in color from dark blue to light green as conditions go from most unstable to neutral, and from light green to red as conditions shift from neutral to most stable.

The last point strongly suggests that, in low friction velocity regimes, the effect of wind direction becomes significantly more relevant than in high u_* regimes. In other words, under stable conditions—where friction velocity tends to be lower—wind direction appears to exert a greater influence on pollutant concentrations compared to neutral or unstable conditions, where higher u_* values dominate the dispersion dynamics. This hypothesis will be further investigated in the final part of the sensitivity analysis, which focuses on the role of wind direction.

4.3.3 Wind direction

The last variable considered in the sensitivity analysis is wind direction. As in the case of u_*, the analysis is conducted on the reduced dataset, where the temperature

dimension has been removed. Consequently, the plot in Fig. 4.8 is obtained by averaging over the set of friction velocities Φ:

$$\overline{C_{\text{mean}}(\varphi)} = \frac{1}{|\Phi|} \sum_{u_* \in \Phi} C_{\text{mean}}(\varphi, u_*, i) \qquad \forall i \in \mathcal{I} \tag{4.5}$$

Since wind direction is a cyclic variable defined in the interval $[0°, 360°]$, the results are represented using polar coordinates. From Fig. 4.8, the following observations can be drawn:

- The previously observed oscillatory behaviour persists: concentration levels decrease when transitioning from unstable to neutral conditions and rise again as conditions become stable.
- As stability increases, the influence of wind direction becomes more pronounced, as evidenced by the progressive flattening of the polar plots. While the profiles under unstable conditions appear nearly circular, stable regimes exhibit more elliptical shapes, indicating stronger directional effects.

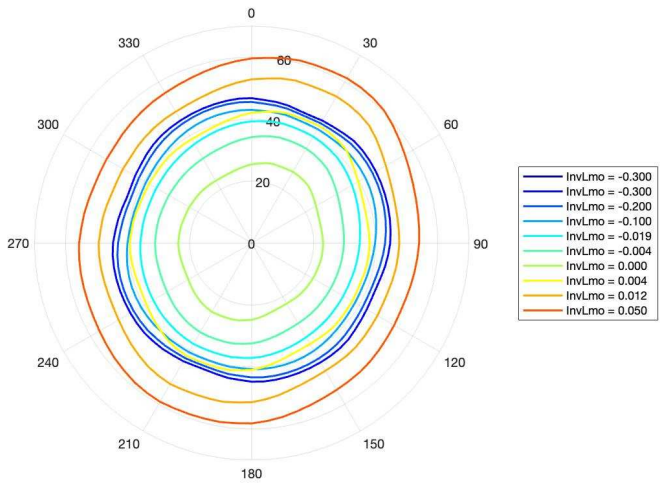


Figure 4.8: Variation of spatially averaged NO_x concentrations as a function of wind direction for different stability regimes. Lines transition in color from dark blue to light green as conditions go from most unstable to neutral, and from light green to red as conditions shift from neutral to most stable.

To further investigate the directional effects and the oscillatory patterns observed, we plotted the concentration profiles as a function of wind direction for each stability class and each feasible value of u_*. From the previous analysis on u_*, we learned that high friction velocities significantly reduce concentration levels. In Fig. 4.9, it

is evident how the combined effect of a shift toward neutral atmospheric conditions and the increase in feasible u_* values contributes to reducing concentration levels. This results in more data points being concentrated near lower concentration values, thereby reducing the average concentration $\overline{C_{\text{mean}}(\varphi)}$ defined in Eq. 4.5.

Furthermore, the directional influence becomes increasingly prominent under stable conditions, as demonstrated by the more elongated profiles. In addition, within the critical region identified in the sensitivity analysis over u_*—specifically in the range [0.05, 0.2]—it was observed that concentration levels are actually higher under unstable conditions compared to stable ones, despite having the same friction velocity. This apparently counterintuitive behaviour can be explained by considering the relative importance of the two dominant transport mechanisms at low wind speeds: turbulent exchange with the overlying atmosphere and longitudinal transport along the street axis.

Under low u_* regimes, turbulent vertical transport plays a more significant role than horizontal advection. Since turbulence is enhanced in unstable atmospheric conditions, the increased vertical exchange tends to retain pollutants within the urban canopy, leading to higher concentration levels compared to stable scenarios, where turbulent exchange is weaker.

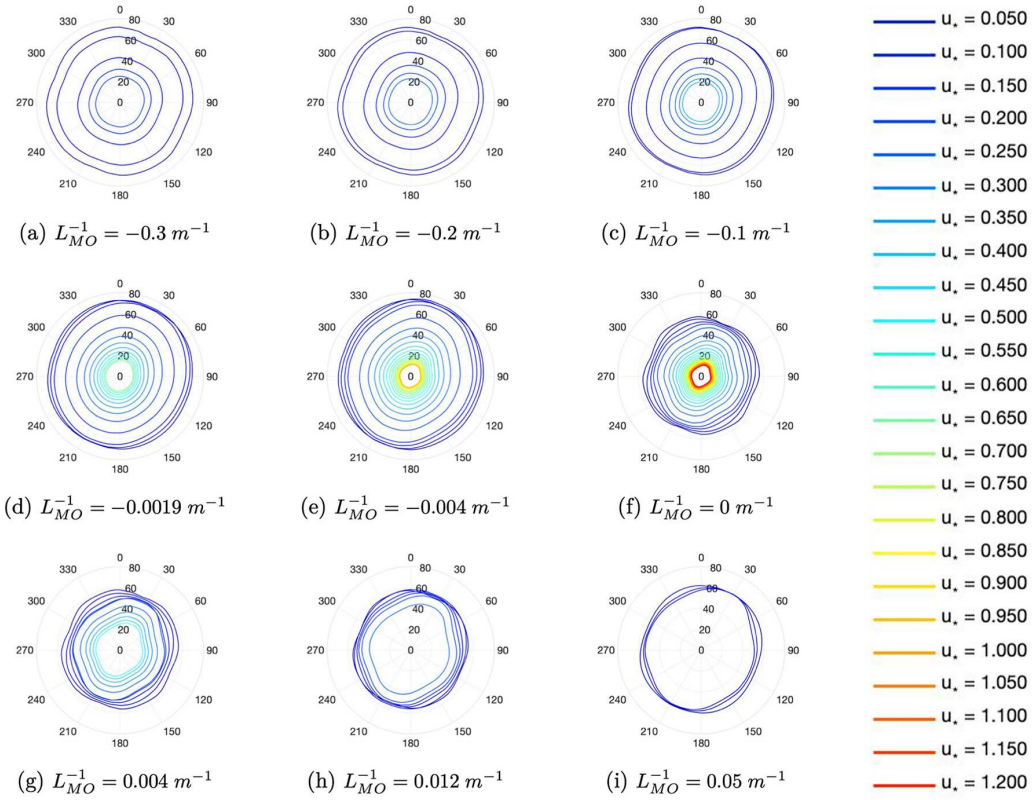


Figure 4.9: Spatially averaged concentrations levels for different values of stability and friction velocity. Panels (a) to (i) represent conditions ranging from the most unstable to the most stable, with panel (f) corresponding to the neutral case. The lines, ranging in color from dark blue to red, correspond to increasing values of u_* from 0.04 m/s to 1.2 m/s.

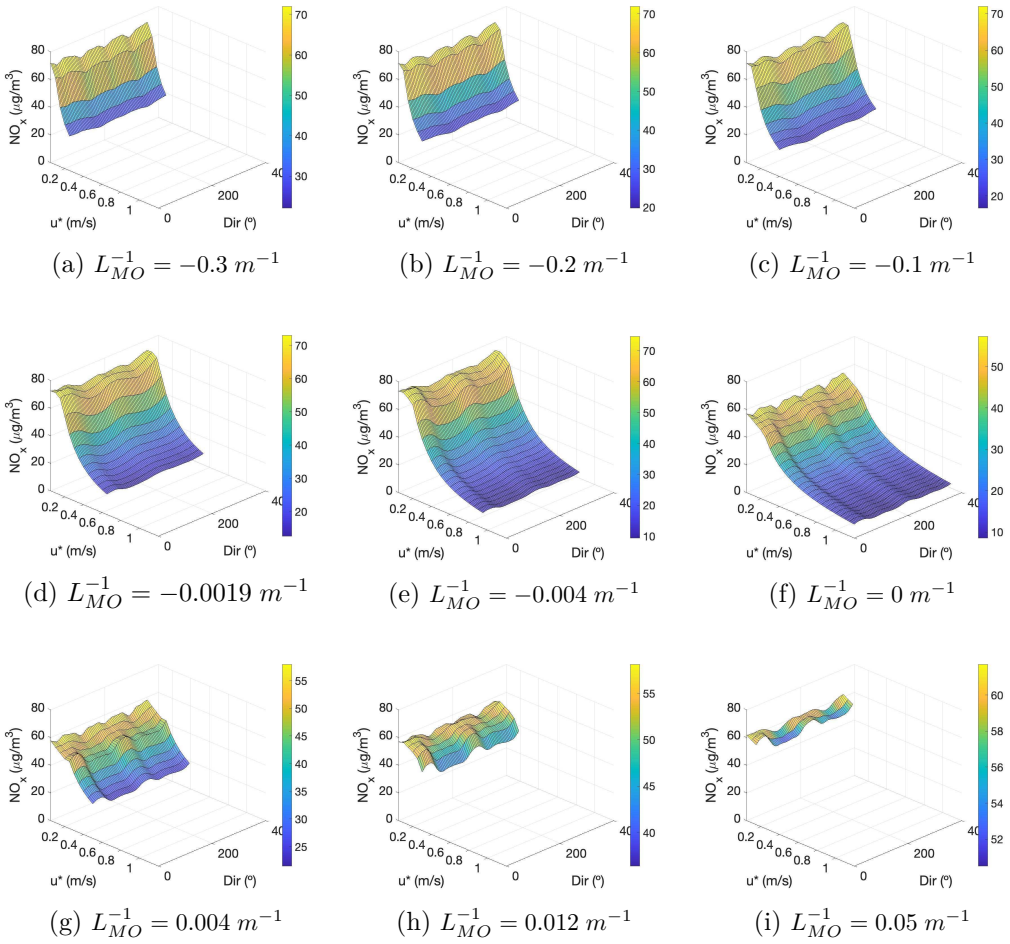
This analysis highlights that, although wind direction becomes increasingly influential in stable regimes, its effect alone is not sufficient to account for the concentration peaks observed under unstable conditions at low friction velocities. Rather, the interplay between stability and vertical turbulent mixing must also be considered to fully interpret this behaviour.

4.4 Summary and implications

The results of the sensitivity analysis, also visually summarized in Fig. 4.10, highlight the dominant role of specific meteorological variables in modulating NO_x concentration levels at street scale. The main takeaways are as follows:

- **Temperature** plays a negligible role in affecting average NO_x concentrations at street level. This allowed us to reduce the dimensionality of the dataset, eliminating the temperature axis and thus improving computational efficiency by a factor of 12.
- **Friction velocity** u_* emerged as the most influential variable due to its dual role:
  1. It controls pollutant advection along the street axis through the term U_H.
  2. It regulates turbulent exchanges with the overlying atmosphere via the vertical component σ_w.

  As u_* increases, concentration levels drop rapidly, particularly in near-neutral conditions where higher values of u_* are more likely.
- **The strong correlation between** L_{MO}^{-1} **and** u_*—a result of boundary-layer physics—implies that high u_* values are only feasible under neutral or quasi-neutral conditions. This:
  1. Causes a natural filtering of meteorologically unfeasible conditions.
  2. Helps explain the lower average concentration levels observed in neutral regimes.
- **Wind direction** becomes increasingly relevant under *stable* conditions. The polar plots showed that directional influence intensifies as stability increases, reshaping the spatial profile from quasi-circular (unstable) to ellipsoidal (stable).
- A **critical regime** was observed in the range $u_* \in [0.05, 0.2]$, where—at fixed u_*—concentration levels are higher in the unstable case than in the stable one. This inversion is likely caused by a stronger contribution of turbulent mixing in unstable conditions compared to street-axis advection.
- **Physical consistency of results**—such as the monotonic decrease of concentrations with u_* and the response to atmospheric stability—confirms the reliability of the synthetic dataset as a basis for training surrogate models and strengthens the methodological soundness of the hybrid physical–statistical approach.
- **Implications for model training:** The sensitivity results will guide feature selection and weighting strategies in the construction of predictive models. In particular, they justify a reduced input space focused on the most informative meteorological drivers.



(a) $L_{MO}^{-1} = -0.3\ m^{-1}$ (b) $L_{MO}^{-1} = -0.2\ m^{-1}$ (c) $L_{MO}^{-1} = -0.1\ m^{-1}$

(d) $L_{MO}^{-1} = -0.0019\ m^{-1}$ (e) $L_{MO}^{-1} = -0.004\ m^{-1}$ (f) $L_{MO}^{-1} = 0\ m^{-1}$

(g) $L_{MO}^{-1} = 0.004\ m^{-1}$ (h) $L_{MO}^{-1} = 0.012\ m^{-1}$ (i) $L_{MO}^{-1} = 0.05\ m^{-1}$

Figure 4.10: Different surfplots representing the spatially averaged concentration levels of NO_x, for all nine different stability regimes, as a function of friction velocity and wind direction. Panels (a) to (i) represent conditions ranging from the most unstable to the most stable, with panel (f) corresponding to the neutral case. The difference in surface area arises from the limited range of feasible u_* values.

Chapter 5

Predictive Model - part I

The sensitivity analysis has demonstrated that pollutant concentrations are strongly dependent on a small set of meteorological parameters. This opens the door to developing data-driven surrogate models capable of rapidly predicting concentrations under arbitrary conditions.

This chapter presents the development of a data-driven predictive model for estimating street-level pollutant concentrations. The first part describes the construction of the training dataset and the methodological choices made during its preparation. The chapter then focuses on interpolation-based approaches, illustrating how these techniques were employed both for dimensionality reduction and to provide valuable insights into the behaviour of the model. Finally, it anticipates how the outcomes of the interpolation analysis informed and supported the regression techniques explored in the following chapter.

5.1 Outline of the work

In Section 3.4 we discussed the role of linear modulation coefficients applied to street emissions, highlighting how these coefficients exhibit significant oscillations over the time window of our simulation. While this aspect is not particularly critical in the context of the sensitivity analysis—where the analysis operates without a temporal frame of reference—it becomes a key challenge when testing predictive models. Indeed, model predictions are validated against SIRANE outputs, which inherently depend on time and therefore on the modulation coefficients applied to emissions. To address this issue, we developed a **two-database strategy**. The first dataset contains concentration levels due exclusively to street emissions (C_{street}), which can thus be modulated post hoc in a linear fashion to match any desired temporal profile. The second dataset contains concentrations resulting solely from surface emissions

(C_{surf}), which are characterized by a constant modulation coefficient and therefore do not require further adjustment. The limitation of this approach lies in the fact that the modulation coefficient must be provided by the user based on historical data, as it cannot be inferred directly from the meteorological inputs themselves. Once these two datasets were generated, we initially investigated the use of linear interpolation as a prediction technique. This allowed us to assess the potential for database size reduction while preserving sufficient accuracy. These results served as a preliminary benchmark before the implementation of more advanced statistical tools, with a particular focus on linear regression methods.

5.2 The database creation

In the creation of the dataset we followed the same approach explored in the first part of Chapter 4: we focused exclusively on street-level NO_x concentration levels, treated as a passive scalar, and neglected point emissions and background concentrations in order to reduce computational costs and accelerate the simulations.

5.2.1 Data structure

As previously discussed, we implemented a two-database strategy and thus generated two five-dimensional arrays, denoted as C_{street} and C_{surf}, both with dimensions:

$$D = 586 \times 73 \times 12 \times 24 \times 9$$

where each dimension respectively corresponds to street index, wind direction, temperature, friction velocity, and inverse Monin-Obukhov length.
The modulation coefficient for street emissions was initially set arbitrarily to unity. To evaluate the correct concentration levels for a given simulation hour t, the C_{street} database is subsequently rescaled using the actual linear modulation coefficient $\alpha(t)$:

$$C^t_{\text{street}} = \alpha(t) \cdot C_{\text{street}} \quad \forall t \in T \tag{5.1}$$

The total concentration in each street is then computed as:

$$C = \alpha(t) \cdot C_{\text{street}} + C_{\text{surf}} + F(t) \quad \forall t \in T \tag{5.2}$$

where $F(t)$ represents the background concentration levels at hour t.
Both $\alpha(t)$ and $F(t)$ are stored in dedicated tables with 8760 rows, corresponding to each simulation hour of the year, and are provided as part of the input data for the

SIRANE annual simulation over the San Salvario district. This data structure relies on the assumption that concentration levels vary linearly with respect to the modulation coefficient. As confirmed in the recent study by Tianyang and Sofia [4], this linearity is a valid hypothesis when dealing with passive scalar diffusion phenomena.

5.3 Concentration function

The datasets C_{street} and C_{surf} contain the street-level NO_x concentrations produced respectively by linear and surface emissions. Both datasets depend exclusively on the meteorological parameters and the street index s, as follows:

$$C_{\text{street}}(s) = f(s, \varphi, T, u, L_{MO}^{-1}), \quad C_{\text{surf}}(s) = f(s, \varphi, T, u, L_{MO}^{-1}) \quad \forall s \in \Sigma \tag{5.3}$$

where Σ denotes the set of streets within the San Salvario district.
By introducing the modulation coefficient $\alpha(t)$ for linear emissions and the background concentration $F(t)$, we extend the structure to incorporate temporal variability. The total concentration for a given street and simulation hour can thus be expressed as:

$$C(s,t) = \alpha(t) \cdot C_{\text{street}}(s) + C_{\text{surf}}(s) + F(t) \quad \forall t \in \Omega, \forall s \in \Sigma \tag{5.4}$$

where Ω represents the set of all simulation hours.
This formulation enables the computation of both spatial and temporal statistics, depending on the variable of interest. It is important to note that the spatial structure is intrinsically embedded in the dataset itself, as SIRANE's outputs are strictly tied to the specific urban geometry provided by the shapefile. As discussed in Chapter 2, the model developed in this work is not general-purpose: it is calibrated specifically for San Salvario and is not readily transferable to other urban contexts without retraining on a new geometry.

5.3.1 Prediction

The predictive models implemented in this work aim to infer concentration levels by learning the relationship between meteorological parameters and pollutant concentrations. The models estimate $C_{\text{street}}(s)$ and $C_{\text{surf}}(s)$ based on the meteorological inputs. Once these estimates are obtained, the total predicted concentration for each street and time step is calculated as:

$$\hat{C}(s,t) = \alpha(t) \cdot \hat{C}\text{street}(s) + \hat{C}\text{surf}(s) + F(t) \quad \forall t \in \Omega, \forall s \in \Sigma \tag{5.5}$$

Both the reference data from the year-long SIRANE simulation and the outputs of the predictive models are stored in matrices of size 586×8760, where rows correspond to street indices and columns to simulation hours. This data structure allows us to easily switch between the temporal and spatial dimensions by simply operating on columns instead of rows, depending on the type of statistic or analysis of interest.

HOURS	0	1	3	...	8759
Id_0	158.62	163.34	170.68	...	80.46
Id_1	157.26	161.72	169.69	...	81.96
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
Id_585	160.79	164.59	170.48	...	88.99

Table 5.1: Example of hourly data for multiple street IDs across a year. The values represents NO_x concentrations levels in $\mu g/m^3$.

5.3.2 Model evaluation metrics

The performance of the predictive models was evaluated using a range of statistical indicators commonly employed in air quality modelling and data science. These metrics were applied both to the spatial dimension (i.e., street-level comparisons at specific time points) and the temporal dimension (i.e., time series at specific locations), depending on the focus of the analysis. The complete set of metrics considered is listed below:

- **Mean Absolute Error (MAE)**:

$$MAE = \frac{1}{N} \sum_{i=0}^{N-1} |y_i - \hat{y}_i|$$

  A robust and interpretable measure of average prediction error in absolute units.

- **Mean Square Error (MSE)**:

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2$$

  Emphasizes larger errors due to the squared term, making it sensitive to outliers.

- **Root Mean Square Error (RMSE)**:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}$$

  Provides an error estimate in the same units as the observed data, facilitating interpretation.

- **Mean Absolute Percentage Error (MAPE)**:

$$MAPE = \frac{100}{N}\sum_{i=0}^{N-1}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

  Useful for understanding error in relative terms; however, it can be sensitive to small denominators.

- **Normalized Mean Square Error (NMSE)**:

$$NMSE = \frac{\overline{(y_i - \hat{y}_i)^2}}{\overline{y_i} \cdot \overline{\hat{y}_i}}$$

  Facilitates the comparison of model performance across different scales of concentration levels.

- **Coefficient of Determination (R^2)**:

$$R^2 = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \overline{y})^2}$$

  Measures the proportion of variance in the observed data explained by the model.

- **Pearson Correlation Coefficient (ρ)**:

$$\rho = \frac{\mathrm{Cov}(y, \hat{y})}{\sigma_y \cdot \sigma_{\hat{y}}}$$

  Quantifies the linear correlation between predictions and observations.

While all these metrics provide valuable insights into different aspects of model performance—such as accuracy, bias, variance explained, and robustness to outliers—for the purpose of comparing different models within this work, the **spatially averaged annual RMSE** was selected as the primary performance indicator. This choice is motivated by several considerations:

- RMSE expresses the error in the same units as pollutant concentrations, making it directly interpretable in terms of air quality standards.
- By averaging spatially over all streets and temporally over a full year, this metric aligns with the type of data typically used by public authorities and environmental agencies to evaluate population exposure, assess compliance with regulations, and formulate policies.
- RMSE, by penalizing larger errors more heavily, ensures that models providing more consistent and reliable predictions across all streets and time periods are favored.

Secondary metrics such as R^2, NMSE, and the correlation coefficient were also examined during model evaluation to provide complementary information on model fit and to identify specific strengths or weaknesses (e.g., systematic bias, ability to capture variability). However, the annual RMSE remains the key metric for benchmarking and reporting purposes in this study.

5.4 Interpolation

The first method tested in this work was simple linear interpolation. Although this technique can produce remarkably consistent predictions with low error, it is not considered a true machine learning approach. Indeed, it does not generate a predictor function trained on a dataset capable of generalising to unseen data. Rather, it always relies on the existing data structure itself to perform interpolation and estimate values. This inherent dependence on the underlying dataset makes the method non-portable, as it cannot be applied to new contexts without the full reference data.
Nonetheless, given its surprisingly high accuracy in our case, linear interpolation was employed as a preliminary tool to explore the feasibility of further reducing the dataset size. Specifically, it was used to investigate whether increasing the discretization step of the remaining meteorological variables (with temperature already excluded) would allow for significant data reduction while maintaining an acceptably low prediction error.

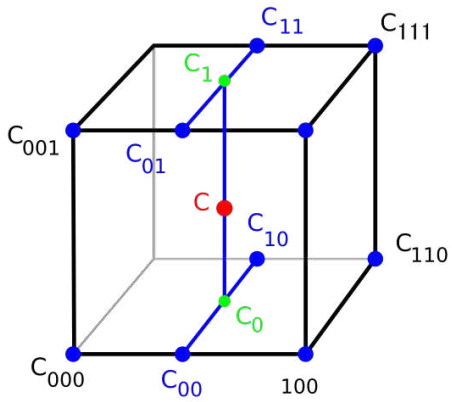


Figure 5.1: Schematic of trilinear interpolation. The function value at point C is computed iteratively from the values at the vertices of the cube that encloses it. The method is based on the progressive reduction of dimensionality: the interpolation is first performed along one axis, then along the second, and finally along the third.

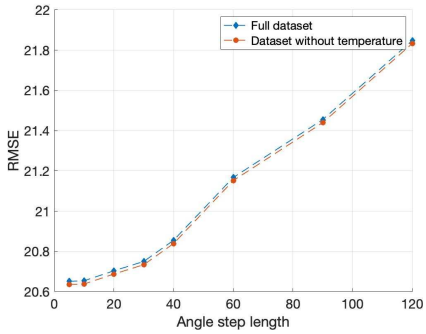
Another important consequence of using linear interpolation lies in its inherent filtering effect on meteorologically feasible conditions. Since the interpolation procedure estimates a value by weighting the contributions of neighbouring data points, it implicitly ensures that predictions are based only on physically plausible combinations of meteorological parameters. In other words, because the target value to be estimated corresponds to an actual, feasible meteorological state, its nearest neighbours in the phase space will also represent realistic atmospheric conditions. This characteristic provides a form of automatic consistency with meteorological constraints, without requiring explicit filtering or additional validation of input combinations.

5.4.1 Dimensionality reduction

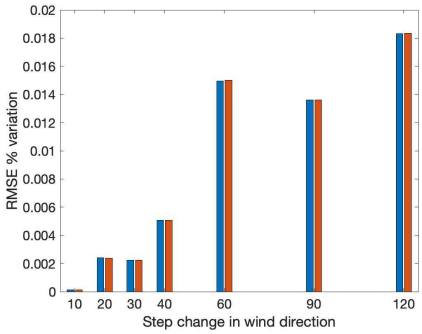
In this section, as anticipated, we investigate how the spatially averaged annual RMSE varies when progressively reducing the resolution of our dataset, i.e., by adopting a coarser discretization for one variable at a time while keeping all other parameters fixed.

Wind Direction

The first variable examined in this analysis is the wind direction. The results are reported in the plots below:



(a) RMSE as a function of the step size for wind direction discretization.



(b) Percentage variation of RMSE between consecutive discretization step sizes.

Figure 5.2: Results of the analysis on wind direction. Panel (a) shows the absolute RMSE trend as the step size of wind direction discretization increases. Panel (b) reports the percentage change in RMSE between successive step size increments. The blue line refers to the interpolation error on the full dataset, while the red line refers to the error on the reduced dataset without temperature. As expected, temperature has no influence on the results.

In the sensitivity analysis, temperature had already been ruled out as a significant variable for influencing concentration levels. Nonetheless, to confirm this conclusion, we performed the interpolation both on the full dataset and on the reduced dataset excluding temperature. The results confirmed that temperature has no effect on interpolation performance. Therefore, from this point onwards, we will definitively exclude temperature from the analysis and focus exclusively on the reduced dataset.

Inverse Monin-Obukhov Length

For this parameter, since it does not belong to a regular interval like the other variables but instead to a set of nine predefined values, we adopted a different approach. Specifically, we constructed an heuristic by selecting and testing various subsets of the original nine elements. The subsets were grouped into three categories, designed to preserve a representative balance of stability conditions:

- **Category 22**: subsets of cardinality 5, including two unstable values, two stable values, and the neutral condition: $(un_1, un_2, 0, st_1, st_2)$;
- **Category 21**: subsets of cardinality 4, including two unstable values, one stable value, and the neutral condition: $(un_1, un_2, 0, st_1)$;
- **Category 11**: subsets of cardinality 3, including one unstable value, one stable value, and the neutral condition: $(un_1, 0, st_1)$.

For each subset in every category, we performed linear interpolation and evaluated the associated RMSE. The results are summarized in Fig. 5.3.

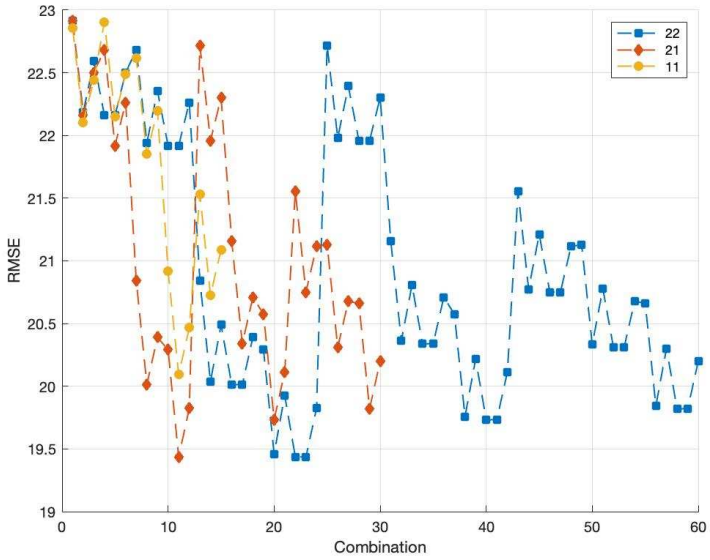


Figure 5.3: RMSE values obtained for different subsets of inverse Monin-Obukhov length values. Blue line: category 22 (5 elements); red line: category 21 (4 elements); yellow line: category 11 (3 elements).

The subset that yielded the lowest RMSE belonged to category 21 and consisted of the following values:

$$L_{MO}^{-1} = \{-0.3,\ -0.004,\ 0,\ 0.012\}.$$

This result is promising, as it suggests that the concentration levels exhibit an approximately linear behavior within both the unstable and stable regimes, with the neutral condition acting as a meaningful transition point. The following boxplot (Fig. 5.4) further support this interpretation, illustrating the variations in spatially averaged concentration as a function of L_{MO}^{-1}.

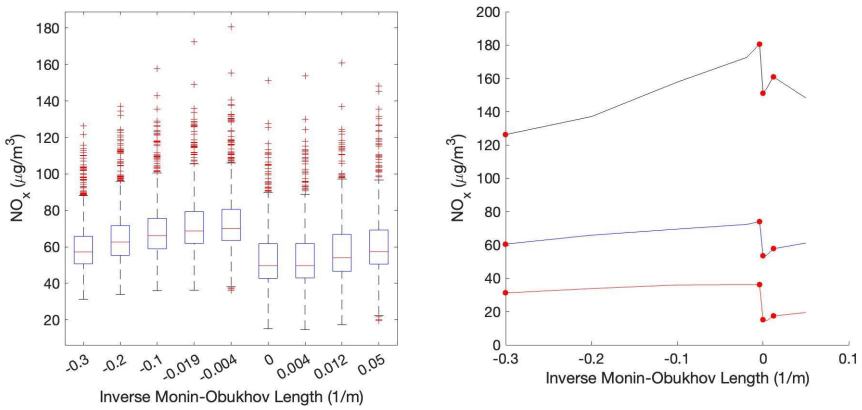
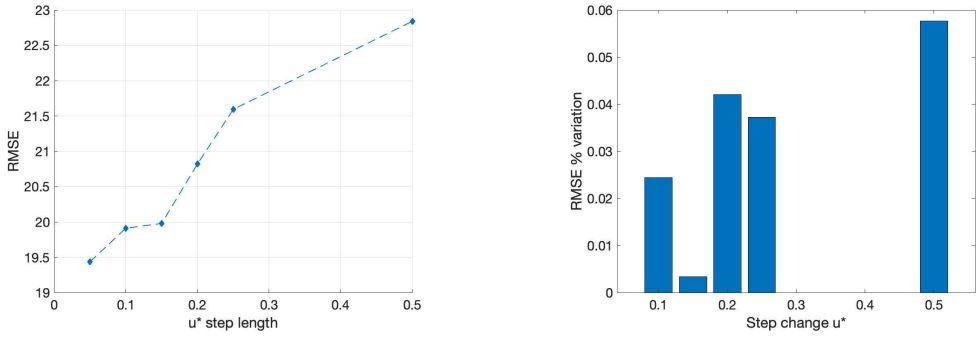


Figure 5.4: Boxplots showing the distribution of NO_x concentrations as a function of inverse Monin-Obukhov length. To the right of the boxplot, a corresponding graph is displayed: the black line represents the trend of maximum concentration values, the blue line represents the average concentration profile, and the red line indicates the minimum concentration as a function of the inverse Monin-Obukhov length.

Friction velocity

The final variable considered in this analysis is the friction velocity u_*. The results are presented in the plots below.



(a) RMSE as a function of the discretization step size for friction velocity.

(b) Percentage variation of RMSE between consecutive discretization step sizes.

Figure 5.5: Results of the analysis on friction velocity. Panel (a) shows the trend of absolute RMSE as the discretization step size for friction velocity increases. Panel (b) reports the percentage change in RMSE between successive step size increments.

From the sensitivity analysis, we have already identified u_* as one of the most influential variables affecting street-level concentration levels. As shown in the plots, even the first coarsening of the discretization step results in an RMSE increase of approximately 2%, confirming its critical role in accurately capturing pollutant dispersion dynamics.

Summary

From this preliminary analysis, the following conclusions emerged:

- Temperature was confirmed to have a negligible influence on street-level NO_x concentrations;
- The RMSE increases only modestly when enlarging the discretization step for wind direction, indicating that this variable is less critical;
- The number of significant values for the inverse Monin-Obukhov length can be reduced to four representative states: the most unstable (-0.3), the least unstable (-0.004), the neutral condition (0), and the most frequently occurring stable condition (0.012);
- The RMSE increases rapidly when the discretization step for friction velocity is enlarged, confirming that this variable is the most influential.

Based on these results, the final decisions for the dimensionality reduction of the phase space were as follows:

1. Removal of the temperature dimension;
2. Increase of the wind direction discretization step from 5° to 30°, which resulted in only a 0.48% increase in RMSE;
3. Reduction of the inverse Monin-Obukhov length values heuristically to a subset of four representative conditions;
4. Maintenance of the original discretization for friction velocity, given its strong influence on concentration levels.

As a result, the total number of meteorological combinations in our synthetic dataset was reduced from 189216 to 1248, corresponding to a reduction of 99.34% compared to the original dataset required by our model. This led to a dramatic reduction in storage requirements: from over 700 GB for the two datasets C_{street} and C_{surf} to approximately 4.6 GB, significantly enhancing the model's portability.
Compared to the one-year simulation over San Salvario, where the number of unique meteorological conditions is 8758 (including temperature) or 8126 (excluding temperature), our reduced synthetic dataset requires only 14.25% and 15.36% of these combinations, respectively.

5.4.2 Testing and results

Once the final dataset was properly reduced, we proceeded with testing the interpolation method using both temporal and spatial metrics. The results are reported graphically and numerically in this section.

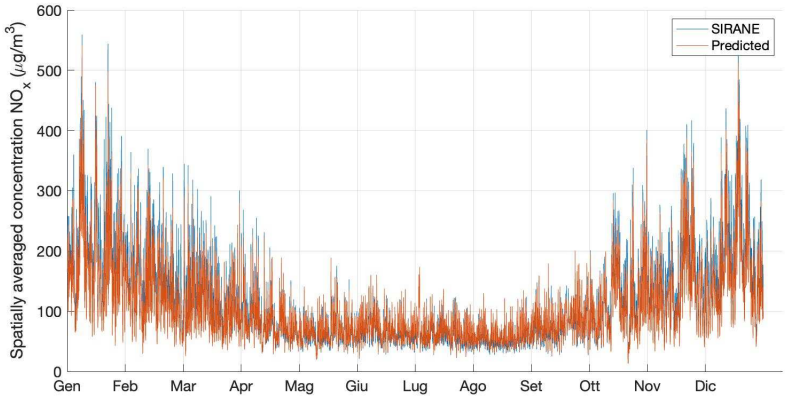
Temporal trends in concentration levels

Table 5.2 summarizes the performance metrics of the interpolation model, evaluated on the annual spatially averaged NO_x concentration levels. The RMSE appears slightly higher than anticipated, likely due to the large temporal window considered (one year), which naturally introduces greater variability. Conversely, the NMSE shows excellent agreement between predictions and reference values. Both the R^2 and the Pearson correlation coefficient confirm a strong correlation, demonstrating that the interpolation model provides reliable estimates of concentration trends.

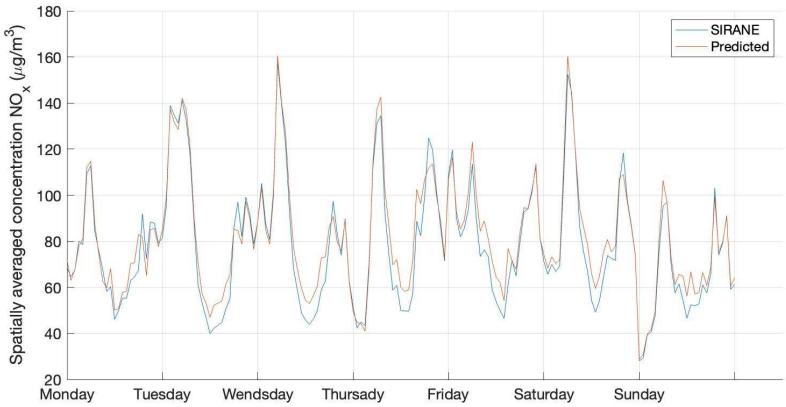
MAE	11.6208
RMSE	19.4791
NMSE	0.0325
MAPE	9.89%
$\mathbf{R^2}$	0.9334
ρ	0.9771

Table 5.2: Performance metrics of the interpolation model evaluated on annual spatially averaged NO_x concentrations.

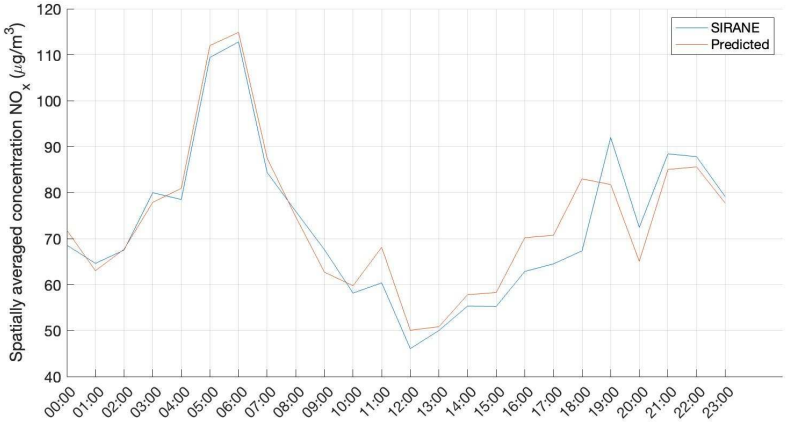
Fig. 5.6 shows the comparison between interpolated predictions and SIRANE reference values for spatially averaged concentrations over different time windows: one year, one week, and one day. The model follows the reference values closely across all time scales. This strong agreement is partly due to the use of the true modulation coefficient as input, which matches the actual temporal modulation applied in the SIRANE simulations. Nonetheless, the interpolation demonstrates good predictive capability in reproducing concentration patterns.



(a) Predicted vs. SIRANE spatially averaged concentrations over one year (2014).



(b) Predicted vs. SIRANE spatially averaged concentrations over one week (07/06/2014 – 14/06/2014).



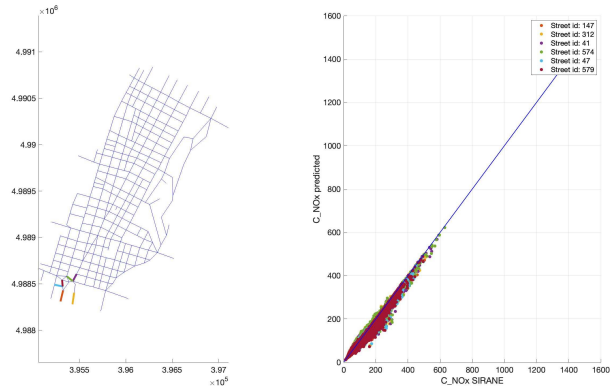
(c) Predicted vs. SIRANE spatially averaged concentrations over one day (07/06/2014).

Figure 5.6: Comparison between interpolated predictions and SIRANE reference values for spatially averaged NO_x concentrations: (a) annual trend, (b) weekly trend, (c) daily trend.

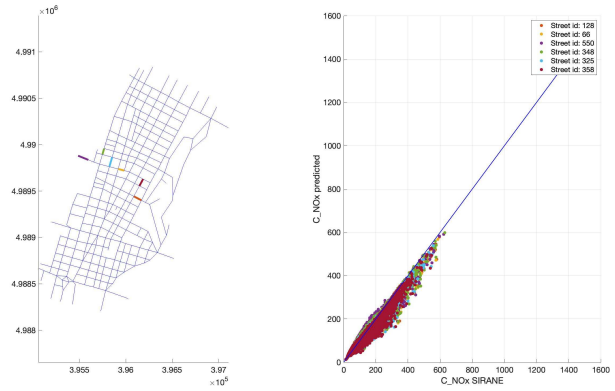
To better understand the spatial variability of model performance, we analyzed the results for three subsets of streets:

- the six streets with the lowest annual RMSE,
- the six streets with RMSE values close to the median,
- the six streets with the highest annual RMSE.

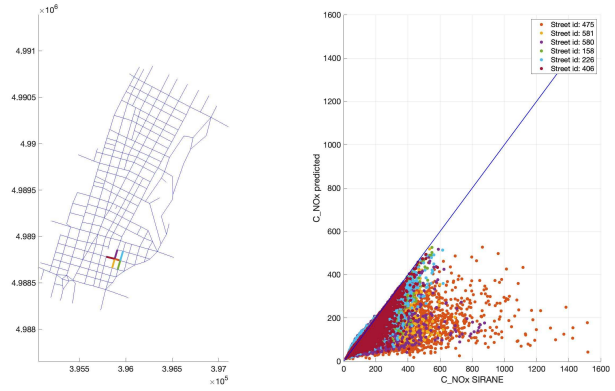
Fig. 5.7 illustrates, for each subset, the comparison between predicted and reference concentrations over one year. It is evident that as the RMSE increases, the model tends to increasingly underestimate concentration levels. Furthermore, streets with the highest RMSE are spatially clustered within a specific area of the district, suggesting possible local effects not fully captured by the interpolation.
Finally, Fig. 5.8 shows the weekly concentration trends for the individual street with the lowest RMSE (ID 147) and the street with the highest RMSE (ID 475). While the model still captures the overall pattern of peaks and troughs, it significantly underestimates peak values in the case of the street with higher RMSE.



(a) Six streets with the lowest annual RMSE: predicted vs. SIRANE concentrations.

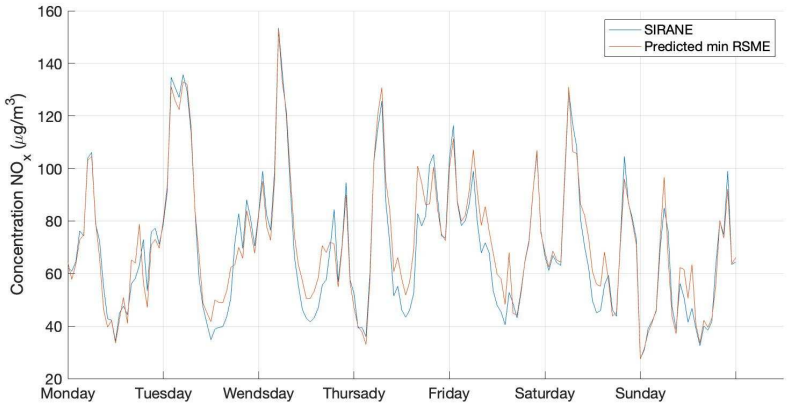


(b) Six streets with median annual RMSE: predicted vs. SIRANE concentrations.

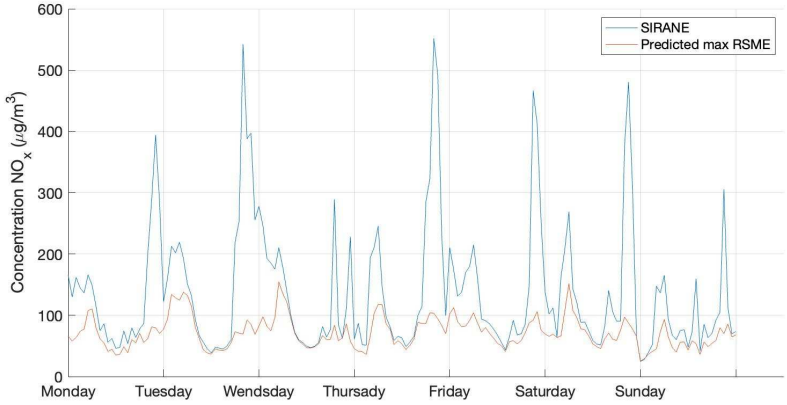


(c) Six streets with the highest annual RMSE: predicted vs. SIRANE concentrations.

Figure 5.7: Comparison between interpolated predictions and SIRANE reference values for NO_x concentrations in selected street subsets.



(a) Predicted vs. SIRANE concentrations over one week for street ID 147 (lowest RMSE).



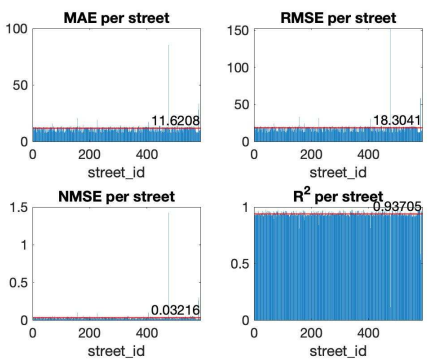
(b) Predicted vs. SIRANE concentrations over one week for street ID 475 (highest RMSE).

Figure 5.8: Weekly comparison between interpolated predictions and SIRANE reference values for two streets with extreme RMSE values.

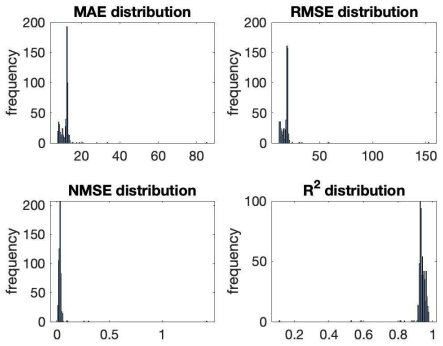
This spatial clustering of high-error streets motivated further investigation through time-averaged RMSE maps to better understand the origin of these discrepancies and assess potential local factors influencing model performance.

Spatial distribution of concentration

In this section, we analyze how the prediction error on time-averaged concentration levels is spatially distributed across the San Salvario district.



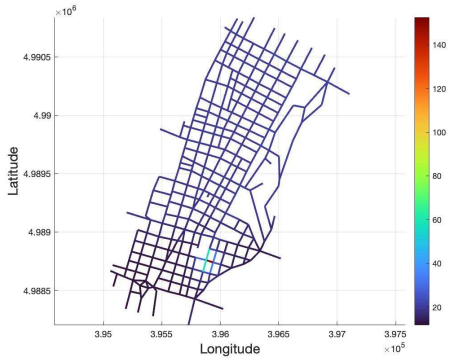
(a) Bar plot of MAE, RMSE, NMSE, R^2 per street. Red line: average value.



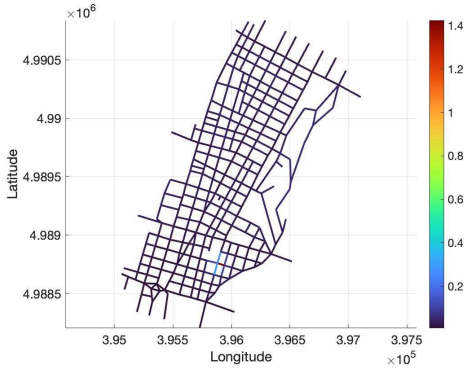
(b) Empirical distribution of MAE, RMSE, NMSE, and R^2 across all streets.

Figure 5.9: Error metrics (MAE, RMSE, NMSE, R^2) for time-averaged NO_x concentration levels: (a) values for each street, (b) corresponding empirical distributions.

From Fig. 5.9, it is clear that while most streets exhibit errors close to the average, a few streets display significantly higher error values, as visible in the spikes in the bar plots. To better understand the spatial pattern of these errors, we mapped the RMSE and NMSE values across the district.



(a) RMSE spatial distribution per street.



(b) NMSE spatial distribution per street.

Figure 5.10: Spatial distribution of RMSE (a) and NMSE (b) for time-averaged NO_x concentrations across San Salvario streets.

Fig. 5.10 highlights how the highest errors are concentrated in a specific area of the district. Comparing this location with the position of point emission sources shown in Fig. 2.6 (see Chapter 2), we observe that the area corresponds precisely to the location of one of the point sources near the entrance of the Michele Lanza underpass. This confirms that the error concentration in that zone is primarily due to the influence of point source emissions. As previously discussed in Section 4.2, these emissions have a limited impact on time-averaged spatial statistics but can

locally affect concentration peaks, which the model—lacking point source contributions—fails to capture.

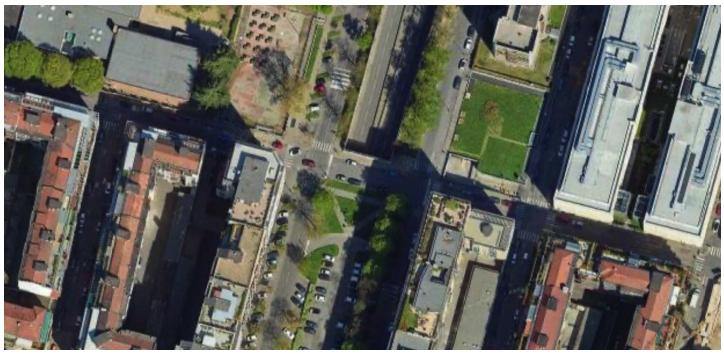
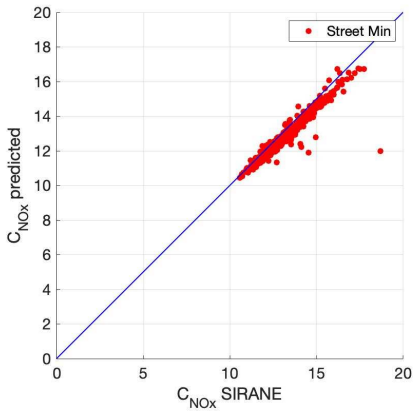
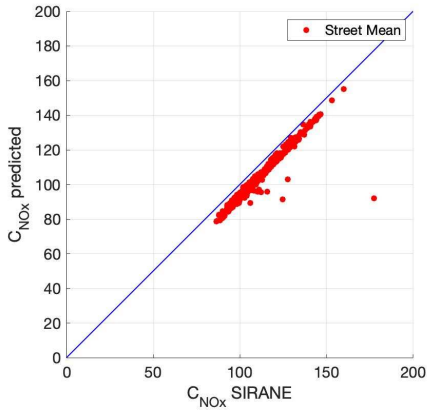


Figure 5.11: Image highlighting the entrance of the Michele Lanza underpass (source: Google Maps).

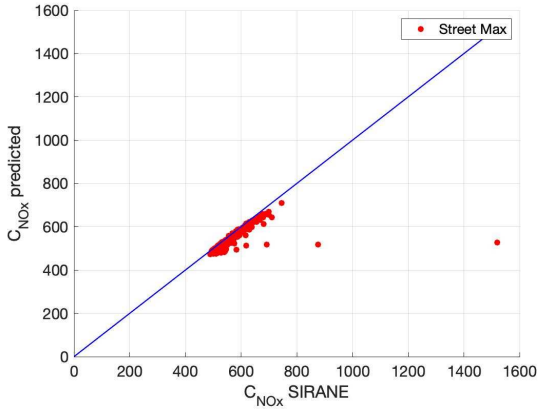
Finally, in the plots of Fig. 5.12, we report how the model predicts the annual minimum, average and maximum NO_x concentration values for each street. The interpolation method shows a tendency to underestimate concentration levels across all three statistics, with a particularly marked underestimation for the annual average and maximum values. A few outliers are also visible in the data. Both the general underestimation and the presence of these outliers can largely be attributed to the absence of point source contributions in the model predictions.



(a) Predicted vs. SIRANE annual minimum NO_x concentrations per street.



(b) Predicted vs. SIRANE annual average NO_x concentrations per street.



(c) Predicted vs. SIRANE annual maximum NO_x concentrations per street.

Figure 5.12: Comparison between predicted and SIRANE reference NO_x concentrations per street for different annual statistics: (a) minimum, (b) average, and (c) maximum values.

5.4.3 Point emission contribution

In this section, we briefly investigate the local effect of different source emissions (street, surface, and point sources). Figure 5.13 shows the contribution of point emissions to the spatially averaged concentration in the San Salvario district over the course of one year. Although their contribution is generally low, there are certain periods in which point sources generate significant concentration peaks that cannot be neglected.

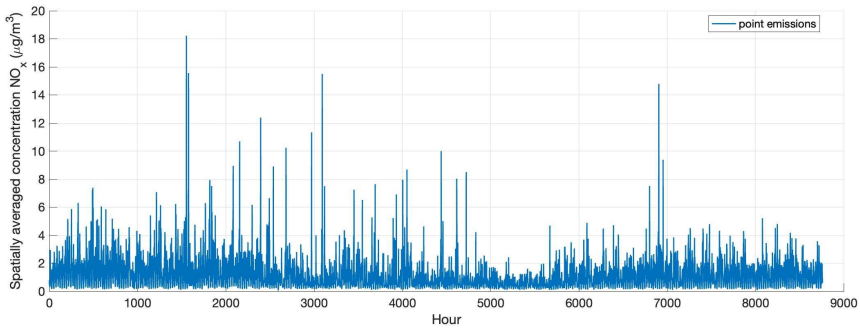
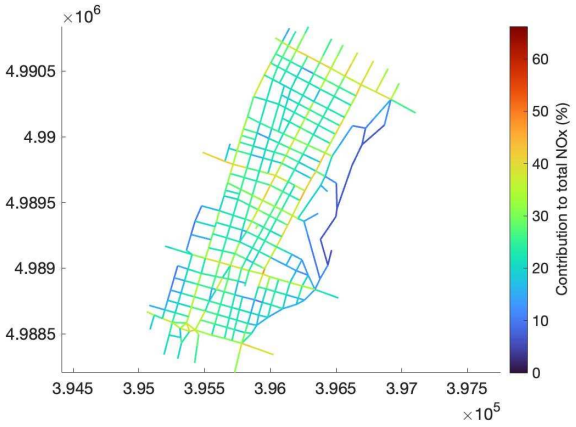
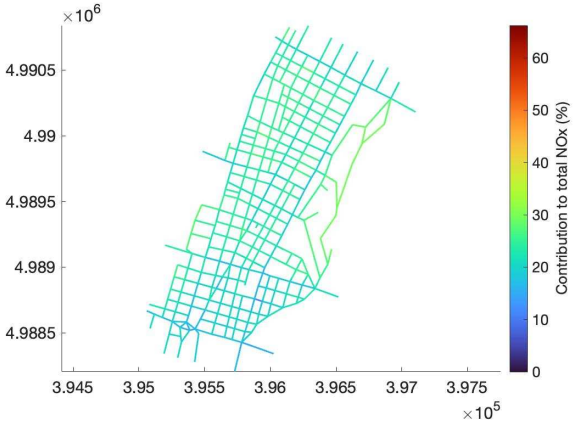


Figure 5.13: Contribution of point emission sources to the spatially averaged NO_x concentration in San Salvario over one year.

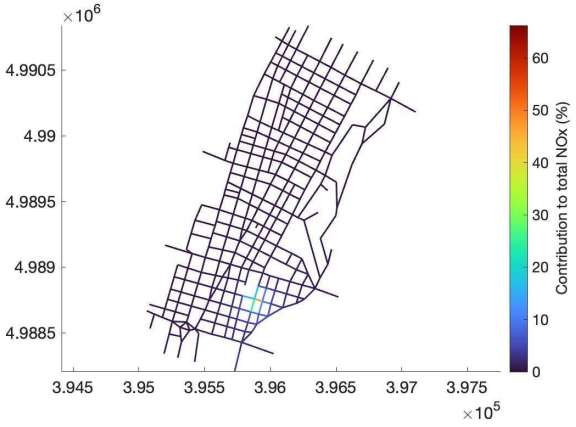
Fig. 5.14 illustrates the spatial distribution of the percentage contribution of the various emission sources (street, surface, and point) to the time-averaged NO_x concentration across the district.



(a) Contribution of street emissions.



(b) Contribution of surface emissions.



(c) Contribution of point emissions.

Figure 5.14: Spatial distribution of the percentage contribution of different emission sources to the time-averaged NO_x concentrations over one year in San Salvario. Panel (a) refers to street emissions, panel (b) to surface emissions, and panel (c) to point emissions.

These results clearly demonstrate that the mismatch in prediction for the specific area corresponding to the maximum RMSE (as seen in street ID 475) is largely attributable to the presence of point source emissions. When adding the point emission contribution to the predicted values for that street, as shown in Fig. 5.15, the agreement with the SIRANE reference significantly improves.

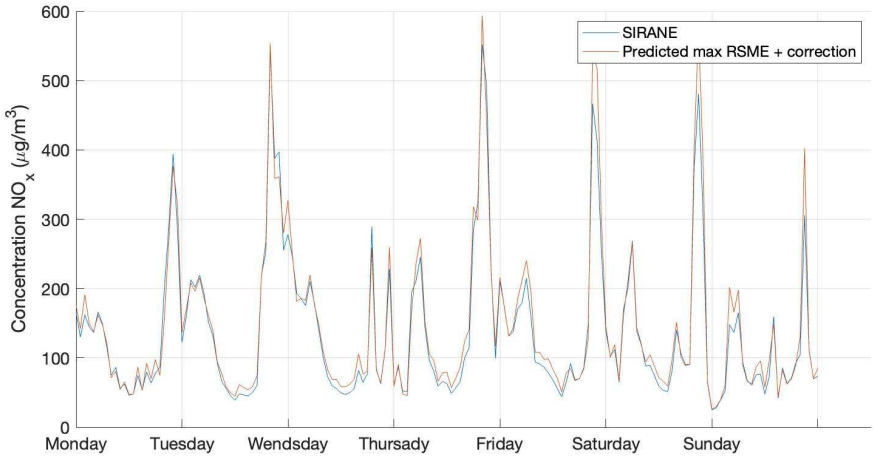


Figure 5.15: Comparison between predicted and SIRANE concentrations for street ID 475 over one week (07/06/2014 – 14/06/2014), after adding the contribution from point emissions as a correction factor.

This finding opens the possibility of adjusting the database by incorporating the hour-by-hour point emission contribution for each street. However, although this approach would reduce the overall error, it is not formally correct: the meteorological conditions used to generate the synthetic database and those associated with the point source emissions are not strictly consistent, and point sources were not part of the learning model. For these reasons, we preferred to evaluate model performance while explicitly acknowledging the bias introduced by neglecting point source contributions.

Chapter 6

Predictive Model – Part II

The interpolation-based method presented in the previous chapter served as a preliminary tool for analysing the dataset and understanding the relationships between meteorological variables and pollutant concentrations. It also provided valuable insights for defining the structure of the predictive model and for guiding dimensionality reduction strategies.

In this final chapter, we extend the predictive framework by introducing regression-based techniques, focusing specifically on linear regression. The chapter describes in detail the setup of the regression models, the rationale behind the choice of predictors, and the implementation strategy. The results will be discussed in continuity with the analyses carried out in the previous chapter, highlighting both the strengths and limitations of the regression approach in reproducing SIRANE's outputs and in supporting data-driven predictions of street-level concentrations.

6.1 Regression

In this section, we describe the linear regression models implemented for predicting street-level NO_x concentrations based on meteorological inputs. Linear regression assumes an additive linear relationship between the predictors (i.e., meteorological variables) and the target variable (pollutant concentration). The general form of the model is:

$$\hat{C} = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon \tag{6.1}$$

where $\hat{C}$ is the predicted concentration, X_i are the predictors, β_i the model coefficients, β_0 the intercept, and ε the residual error term, assumed to follow a normal distribution with zero mean.

This method offers simplicity, computational efficiency, and interpretability, although it may not capture complex non-linear relationships present in the data.

6.2 Predictors and model structure

In designing the regression models for the prediction of street-level NO_x concentrations, two distinct configurations were implemented to balance model simplicity, accuracy, and generalizability. Both models rely on meteorological predictors that influence pollutant dispersion and concentration within the urban environment.

6.2.1 Choice of predictors

The selected meteorological variables were:

- **Wind direction** (φ): represented through its sine and cosine components to handle its circular nature and ensure continuity between 0° and 360°;
- **Friction velocity** (u_*): a key driver of both advection along street axes and turbulent exchange between street canyons and the overlying atmosphere;
- **Inverse Monin-Obukhov length** (L_{MO}^{-1}): an indicator of atmospheric stability, influencing vertical mixing and dispersion processes.

These variables were selected based on their physical relevance and their established role in regulating pollutant transport, as highlighted in the sensitivity analysis presented in Chapter 4.

6.2.2 Model configurations

To explore the trade-off between generalizability and predictive accuracy, two linear regression configurations were tested:

1. **Model 1:** $NO_x \sim \sin(\varphi) + \cos(\varphi) + u_* + L_{MO}^{-1}$ This model is compact and independent of street-specific identifiers. It aims to provide a general predictor applicable to new or unseen streets and different urban contexts, without requiring additional spatial information at prediction time.
2. **Model 2:** $NO_x \sim \sin(\varphi) + \cos(\varphi) + u_* + L_{MO}^{-1} + \text{Road\_ID}$
   This model includes a categorical predictor for the street (Road_ID), allowing it to capture fixed structural differences between streets, such as geometry, orientation, and surrounding morphology. It is designed for greater accuracy

at the individual street level but requires knowledge of the specific street at prediction time and is not generalizable beyond the streets present in the training dataset.

The first model offers scalability and applicability across different scenarios, but at the cost of reduced accuracy in capturing local heterogeneities. The second model achieves a better fit to the observed data but increases complexity proportionally to the number of streets and sacrifices generalizability to new environments.
All models were trained on the interpolation-reduced datasets developed in the previous chapter, constrained only to meteorologically feasible conditions as defined in Chapter 4. The performance of the models was then evaluated against the reference dataset provided by the one-year SIRANE simulation over the San Salvario district.

6.2.3 Testing and results

In this section, we present the evaluation of the two regression models using the same performance metrics introduced in the previous chapter. The goal is to compare their predictive capability, both in temporal trends and in spatial distribution.

Analysis of regression coefficients

Before discussing the predictive performance, it is useful to briefly analyze the regression coefficients obtained for the two models, which are identical for both Model1 (without Road_ID) and Model2 (with Road_ID) as they refer to the same underlying physical relationships. The coefficients were estimated separately for the datasets C_{street} and C_{surf}.
For the C_{street} dataset, the model yielded:

Predictor	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	45.417	0.497	91.466	0
Sin(φ)	0.114	0.030	3.793	1.49e-4
Cos(φ)	0.367	0.028	13.101	3.30e-39
u_*	-42.776	0.065	-662.35	0
L_{MO}^{-1}	65.741	0.348	188.66	0

Table 6.1: Regression coefficients for the model trained on C_{street}.

- A positive intercept (45.417), representing the baseline concentration level in the absence of meteorological influences.
- Small positive contributions from sin(φ) (0.11422) and cos(φ) (0.36734), confirming the limited but non-negligible role of wind direction. The positive sign

indicates that, on average, higher concentrations are associated with wind components blowing along certain preferential directions (depending on the geometry of San Salvario), particularly those aligned with positive sine and cosine projections.

- A strong negative coefficient for friction velocity (u_*: -42.776), consistent with the physical expectation that higher turbulence enhances pollutant dispersion and lowers concentration levels.

- A large positive coefficient for inverse Monin-Obukhov length (L_{MO}^{-1}: 65.741), highlighting how stable conditions (higher L_{MO}^{-1}) hinder vertical mixing and increase ground-level concentrations.

For the C_{surf} dataset, the coefficients exhibit similar trends but differ in magnitude:

Predictor	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	19.169	0.188	101.95	0
Sin(φ)	-0.322	0.011	-28.249	2.26e-175
Cos(φ)	0.565	0.011	53.192	0
u_*	-17.379	0.024	-710.67	0
L_{MO}^{-1}	10.228	0.132	77.512	0

Table 6.2: Regression coefficients for the model trained on C_{surf}.

- A smaller intercept (19.169), in line with the typically lower contribution of surface emissions to overall concentrations.

- A negative coefficient for $\sin(\varphi)$ (-0.32215) and a stronger positive coefficient for $\cos(\varphi)$ (0.56473), suggesting a stronger directional dependency for surface emissions, with concentrations tending to decrease or increase depending on the alignment of wind direction with street orientation. The sign reflects the geometric alignment of the urban grid relative to the cardinal directions encoded by the sine and cosine terms.

- A negative coefficient for friction velocity (-17.379), confirming again the dispersive role of turbulence, though with a reduced impact compared to street emissions.

- A positive coefficient for L_{MO}^{-1} (10.228), showing a smaller but still relevant influence of atmospheric stability on surface-emission concentrations.

All coefficients are statistically significant (p-values close to zero), demonstrating that the selected meteorological predictors provide meaningful explanatory power

for both types of emissions. The sign and magnitude of the coefficients are consistent with the findings of the sensitivity analysis presented in Chapter 4, and they reflect the interaction between wind direction, atmospheric stability, turbulence, and the urban geometry of San Salvario.

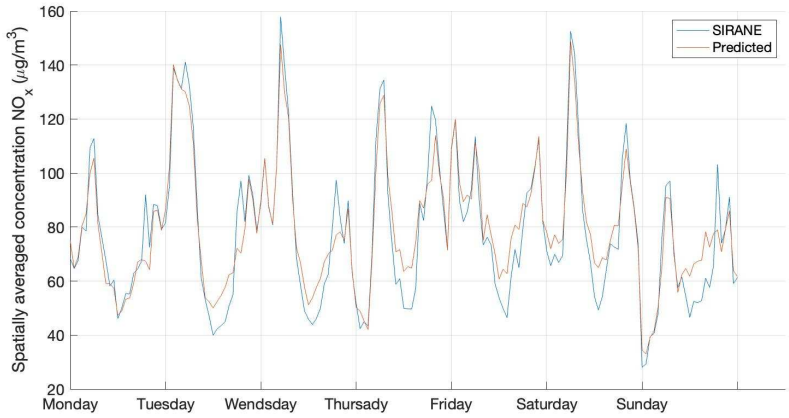
Temporal trends in concentration levels

Table 6.3 summarizes the performance metrics of the two models, evaluated on the annual spatially averaged NO_x concentration levels. As expected, both regression models perform slightly worse than the interpolation-based approach, reflecting the increased challenge of fitting a general predictive function. However, the model including `Road_ID` (Model 2) shows better accuracy across all error metrics, although the improvement over Model 1 is moderate.

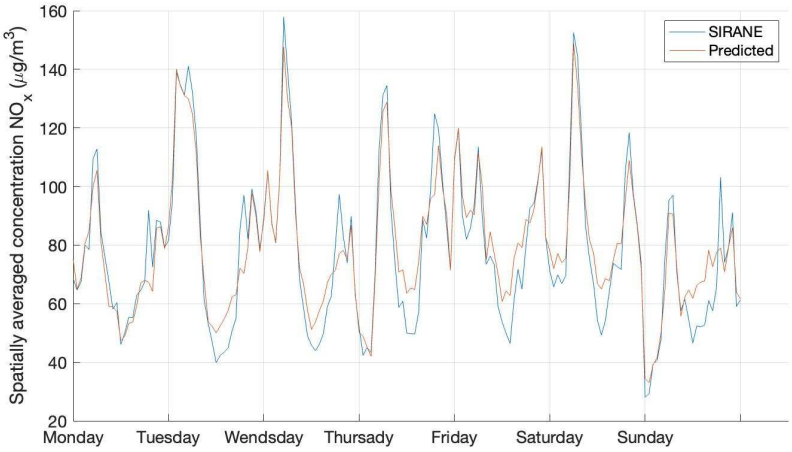
Metric	**Model 1 (No Road_ID)**	**Model 2 (Road_ID)**
MAE	18.2568	16.1574
RMSE	29.3799	26.4485
NMSE	0.0762	0.0617
MAPE	16.60%	13.97%
$\mathbf{R^2}$	0.8485	0.8772
ρ	0.9411	0.9572

Table 6.3: Comparison of error metrics for the two models: Model 1 without Road_ID and Model 2 with Road_ID.

Both models demonstrate a high correlation with SIRANE reference values, as illustrated in Fig. 6.1, where the predicted spatially averaged concentrations over a representative week align closely with the simulated data and the outputs of the two models are visually indistinguishable.



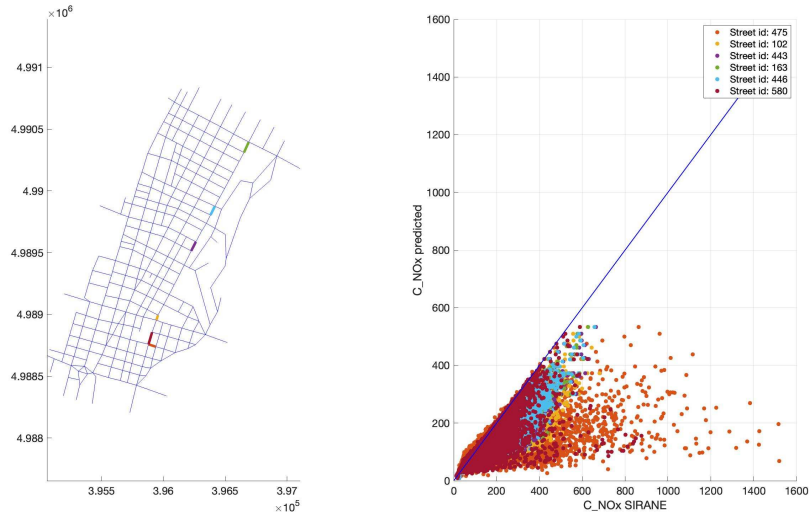
(a) Predicted vs. SIRANE spatially averaged concentrations over one week (07/06/2014–14/06/2014), Model 1 (No Road_ID).



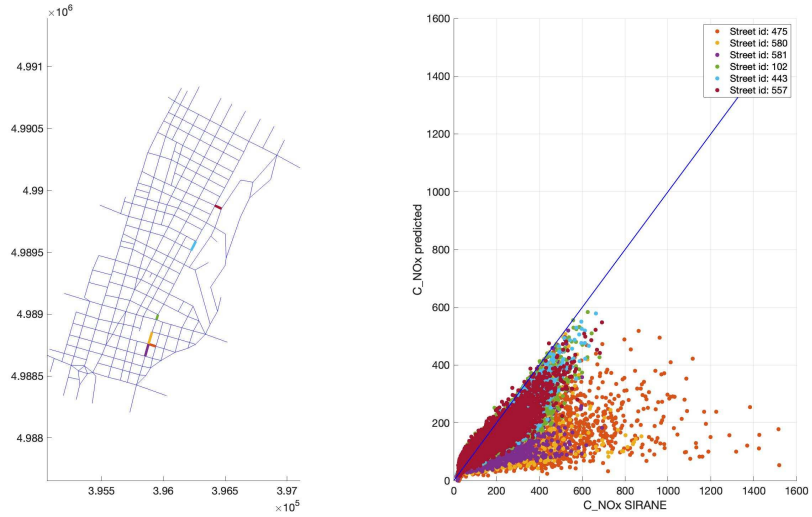
(b) Predicted vs. SIRANE spatially averaged concentrations over one week (07/06/2014–14/06/2014), Model 2 (Road_ID).

Figure 6.1: Weekly comparison between regression predictions and SIRANE reference values for spatially averaged NO_x concentrations: (a) Model 1 (No Road_ID), (b) Model 2 (Road_ID).

The performance on the six streets with the highest RMSE confirms that Model 2 offers greater spatial precision, particularly in areas influenced by local emission sources (e.g., near point sources identified in the interpolation analysis). This is illustrated in Fig. 6.2.



(a) NO_x concentrations in the six streets with the highest RMSE, Model 1 (No Road_ID).

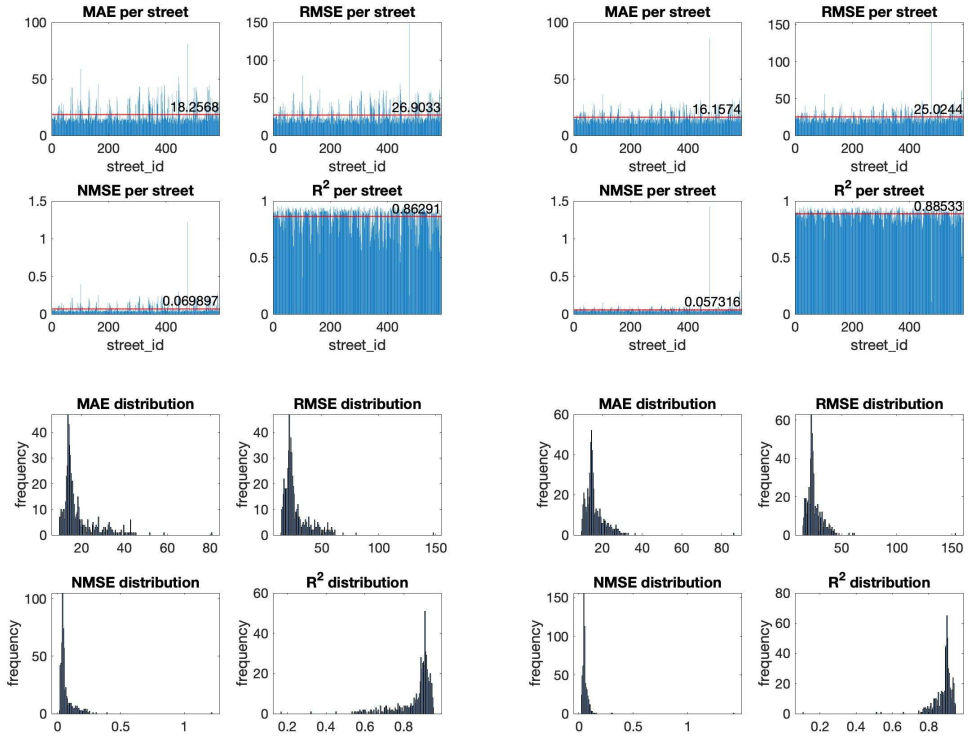


(b) NO_x concentrations in the six streets with the highest RMSE, Model 2 (Road_ID).

Figure 6.2: Comparison of model predictions with SIRANE reference values for NO_x concentrations in the six streets with the highest RMSE.

Spatial distribution of concentration

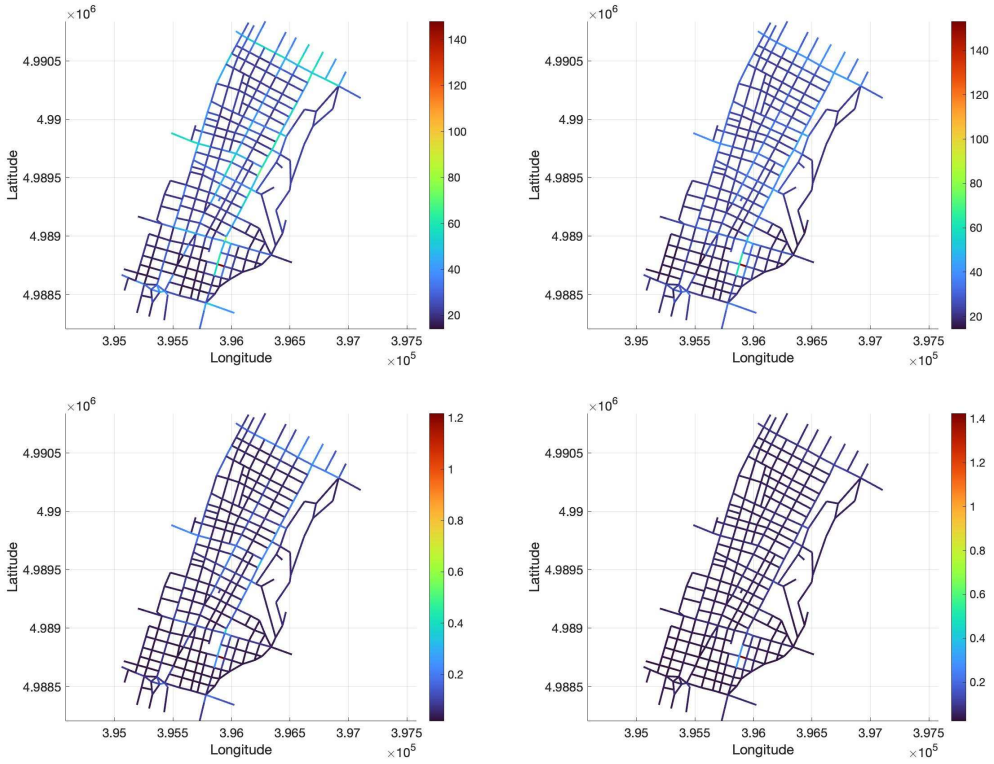
Figure 6.3 presents the distribution of error metrics across all streets. Model 2 achieves slightly lower average error and reduced variance, demonstrating better stability in predictions.



(a) Model 1 (No Road_ID). (b) Model 2 (Road_ID).

Figure 6.3: Error metrics (MAE, RMSE, NMSE, R^2) and their empirical distribution across all streets: (a) Model 1 (No Road_ID), (b) Model 2 (Road_ID). Red lines indicate the average value.

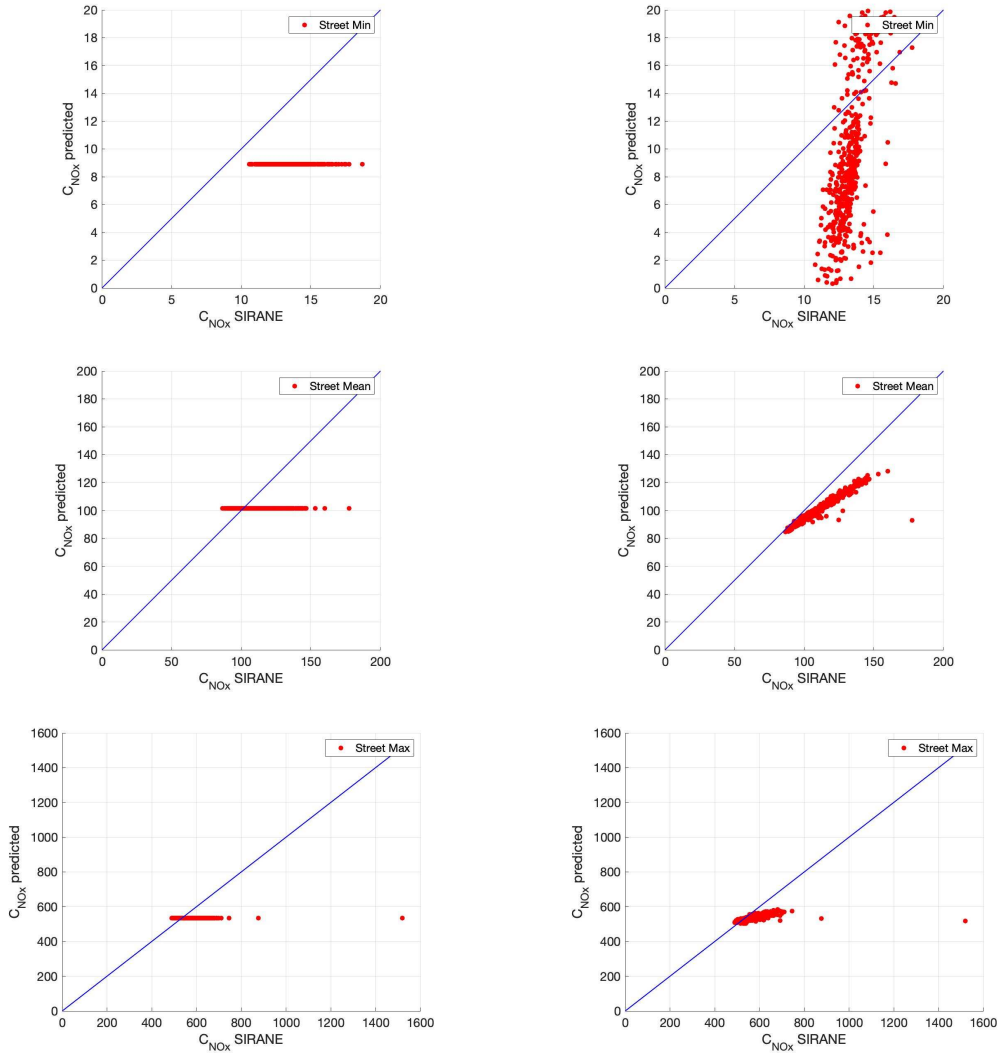
Figure 6.4 maps the spatial distribution of RMSE and NMSE. Model 2 displays more homogeneous performance across the district, especially for NMSE, with errors primarily concentrated around known hotspots, such as the Lanza underpass area linked to point source emissions.



(a) Model 1 (No Road_ID). (b) Model 2 (Road_ID).

Figure 6.4: Spatial distribution of RMSE and NMSE per street: (a) Model 1 (No Road_ID), (b) Model 2 (Road_ID).

Finally, Fig. 6.5 compares how the two models predict annual minimum, mean, and maximum concentrations across streets. Model 1 tends to average out street-specific variability since it is not able to learn that the differences in the concentrations levels are due to urban geometry. Model 2 is able to distinguish between streets but still shows an overall tendency to underestimate concentrations. The worst performance is observed in the annual minimum predictions, with high variance in the results and scattered data points.



(a) Model 1 (No Road_ID). (b) Model 2 (Road_ID).

Figure 6.5: Comparison of annual minimum, mean, and maximum NO_x concentrations across streets for the two models (a) Model 1 (No Road_ID), (b) Model 2 (Road_ID).

6.3 Final considerations

The most accurate model performance was achieved using interpolation. This is not unusual, as interpolation strongly relies on the dataset structure itself to estimate values and inherently incorporates the geometry of the district, given that the street index is encoded as the first dimension of the data structure.
Regarding the regression techniques, Model 2 — which includes Road_ID as a categorical predictor — provided better performance. However, the improvement offered by this model is not substantial enough to justify its adoption, especially considering that street-level predictions remain less accurate. Furthermore, its complexity scales with the number of streets, and applying this model to a full city domain (where the number of streets can reach the order of 10^4) would result in significant computational expense.
Model 1 — without Road_ID as a categorical variable — performed surprisingly well. This can be attributed to the fact that, as with interpolation, the geometry of the district is implicitly encoded in the dataset. While this model fails to capture accurate local variability, its overall performance remains comparable to that of Model 2. It is important to note, however, that when applied to a different domain, Model 1 would likely fail to provide reliable predictions, as it depends heavily on the geometry-encoded structure of the dataset. Nevertheless, in cases where a complete dataset is available — as in our study — this model represents a preferable choice due to its simplicity, reduced computational requirements, and ease of training.

Conclusions

The work presented in this thesis explored the development of data-driven surrogate models for the prediction of street-level NO_x concentrations in an urban district, using San Salvario (Turin) as a case study. The approach combined physical simulations performed with SIRANE with statistical and regression techniques applied to a synthetic dataset of meteorological scenarios.

The main advantage of the surrogate models lies in their ability to reproduce the concentration levels estimated by SIRANE with significantly reduced computational effort. In particular:

- The interpolation-based model demonstrated high predictive accuracy, thanks to its direct reliance on the structured dataset that encodes the geometry of the urban environment. This approach is efficient and easy to implement when a dense reference dataset is available.
- The linear regression models, especially the configuration including Road_ID as a categorical predictor, allowed the capture of structural differences between streets and improved local accuracy. The simpler model without Road_ID showed a good balance between accuracy and portability, requiring fewer computational resources and offering the potential for application in other urban contexts, provided an adequate dataset is available.

Despite these advantages, the surrogate models developed in this work present important limitations:

- The models are strongly overfitted to the specific urban grid of San Salvario. A truly general-purpose model would need to learn the relationship between urban geometry characteristics, meteorological variables, and street-level concentrations. This could be achieved without necessarily resorting to highly complex architectures (such as convolutional neural networks or graph neural networks), but rather by introducing additional predictors that encode urban

structure — for example, by clustering streets according to their orientation, width-to-length ratio (W/L), or width-to-height ratio (W/H), and predicting concentrations at the level of these classes.

- The temporal dependency in the current models is introduced solely through the user-given modulation coefficients of street emissions. A more advanced model could learn this modulation directly from historical time series data, enabling it to predict temporal emission patterns as a function of meteorological conditions and other contextual variables.
- The time available for this thesis was not sufficient to implement and tune more complex models, such as artificial neural networks or ensemble machine learning techniques, which would have required significant effort both in terms of model architecture design and computational resources for training.

In conclusion, this work represents a preliminary stage of a broader modelling effort. The surrogate models developed provide a valuable and computationally efficient tool for estimating pollutant concentrations in a specific domain, but future work should focus on increasing generalizability, integrating additional urban and temporal features, and exploring more sophisticated learning algorithms. With more time and resources, this approach could be extended to larger urban areas, incorporate richer datasets, and support advanced applications in air quality management and urban planning.

Computational resources provided by hpc@polito, which is a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino (http://www.hpc.polito.it)

Bibliography

[1] M. Bo. *Study of aerosols air pollution assessments in indoor and outdoor environments based on measuring and modelling approaches.* Ph.D. Thesis, 2022.

[2] M. Bo, C.V. Perrine, M. Clerico, C.V Nguyen, F. Pognant, L. Soulhac, and P. Salizzoni. Urban air quality and meteorology on opposite sides of the alps: The lyon and torino case studies. *Urban Climate 34*, page 100698, 2020.

[3] J.C. Chang and S.R. Hanna. Air quality model performance evaluation. *Meteorol. Atmos. Phys. 87*, pages 167–196, 2004.

[4] T. Li, S. Fellini, and R. Van Maarten. Urban air quality: What is the optimal place to reduce transport emissions? *Atmospheric Environment 292*, page 119432, 2023.

[5] T.R. Oke, G. Mills, A. Christen, and J.A. Voogt. *Urban Climates.* Cambridge University Press, 2017.

[6] L. Soulhac, V. Garbero, P. Salizzoni, P. Mejean, and R. Perkins. Flow and dispersion in street intersections. *Atmospheric Environment 43*, pages 2981–2996, 2009.

[7] L. Soulhac, C.V Nguyen, P. Volta, and P. Salizzoni. The model sirane for atmospheric urban pollutant dispersion. part iii: Validation against no2 yearly concentration measurements in a large urban agglomeration. *Atmospheric Environment*, pages 377–388, 2017.

[8] L. Soulhac, R.J. Perkins, and P. Salizzoni. Flow in a street canyon for any external wind direction. *Boundary-Layer Meteorology 126 (1)*, pages 365–388, 2008.

[9] L. Soulhac and Perkins R.J. A new model for ow and dispersion in a street-canyon. *Air Pollution Modelling and Its Application 13*, pages 475–483, 2000.

[10] L. Soulhac, P. Salizzoni, F.-X. Cierco, and R.J Perkins. The model sirane for atmospheric urban pollutant dispersion; part i, presentation of the model. *Atmospheric Environment 45*, pages 7379–7395, 2011.

[11] L. Soulhac, P. Salizzoni, P. Mejeana, D. Didier, and I. Rios. The model sirane for atmospheric urban pollutant dispersion; part ii, validation of the model on a real case study. *Atmospheric Environment 49*, pages 320–337, 2012.