UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN TECNOLOGIE PER LA SALUTE, BIOINGEGNERIA E BIOINFORMATICA
XXXIII CICLO - 2020

SPATIAL ENABLEMENT AND SIMULATION TOOLS TO IMPROVE PUBLIC HEALTH AND WELLBEING IN BIG CITIES: A NEW FRAMEWORK BASED ON THE EUROPEAN PULSE PROJECT

PhD Thesis by
DANIELE PALA

Advisor:
Prof. Riccardo Bellazzi

PhD Program Chair:
Prof. Silvana Quaglini

Acknowledgments

I owe much gratitude to my supervisor Prof. Riccardo Bellazzi and my tutor Prof. Vittorio Casella, who allowed me to take part to this interesting project and guided me through an uplifting research and life experience.
I am very thankful also to Prof. José Pagàn, from the New York University, and Lisa, Kumbie, Yan and Foram from the New York Academy of Medicine, who gave me the possibility to live an incredible experience in New York City that added a lot to my personal and academic life.

I gratefully acknowledge all my colleagues, in particular Giovanna, Nicola and Laura, and all the other people in the BMS, BMI and Geomatics lab, who I shared a lot of coffee breaks and laughter with.

Abstract (Italiano)

Secondo l'Organizzazione Mondiale della Sanità, circa metà della popolazione mondiale vive nelle grandi città del pianeta, e questa proporzione è destinata ad aumentare nei prossimi decenni. Questo cambiamento demografico, insieme al sempre più rapido progresso tecnologico, sta portando a una variazione nella quantità e tipologia di fattori ambientali ai quali siamo esposti nel corso della nostra vita, nonché a un cambiamento dei nostri comportamenti e dello stile di vita. Diversi studi hanno dimostrato che è in corso un aumento della prevalenza di diverse malattie respiratorie e cardiovascolari nella maggior parte del mondo urbanizzato, e alcuni hanno anche ipotizzato che la causa potesse essere identificata in questi cambiamenti. Per facilitare lo studio di questi fenomeni, stanno rapidamente emergendo nuovi campi di ricerca, come per esempio l'*esposomica*, ovvero lo studio dell'effetto di tutti i fattori interni ed esterni a cui siamo esposti che possono influenzare la nostra salute. L'esposomica può essere vista come parte del più grande ambito della *Public Health* (salute pubblica), la scienza che si occupa della prevenzione delle patologie e del miglioramento del benessere basandosi su uno sforzo congiunto di società pubbliche e private e singoli individui.

In questo contesto, negli ultimi anni sono stati sviluppati diversi progetti di salute pubblica, basati sull'utilizzo delle tecnologie più innovative per il miglioramento delle strategie di prevenzione sanitaria. Uno di questi è il progetto PULSE, finanziato dalla Commissione Europea nel 2016, con lo scopo di creare un sistema collaborativo multi-tecnologico per la prevenzione di asma, diabete di tipo 2 e malattie cardiovascolari nelle grandi città. La scelta di limitare lo studio alle grandi città non è casuale: le grandi città sono ambienti fortemente eterogenei, in cui le differenze demografiche, ambientali e socioeconomiche sono spesso pronunciate, e i fattori di esposizione possono variare in modo

significativo anche tra aree geografiche molto vicine tra loro. Sono inoltre diversi gli studi che hanno dimostrato che le disuguaglianze sociali, l'inquinamento atmosferico e lo stile di vita movimentato delle grandi città portano a un rischio di sviluppare alcune patologie, tra cui quelle studiate all'interno di PULSE, più elevato rispetto alle aree non urbanizzate. Lo scopo di PULSE è stato quindi quello di costruire un insieme di strumenti per favorire la comprensione di questi fenomeni e facilitare l'ideazione di interventi volti a migliorare la salute e il benessere. A questo scopo, i sistemi sviluppati in PULSE includono varie componenti tecnologiche che rendono accessibili i dati relativi alla città sia ai cittadini che alle autorità che possono prendere decisioni in campo sanitario. Un'attenzione particolare è data alla dimensione spaziale, in quanto i dati di salute pubblica non possono prescindere dalla componente geografica che li caratterizza nello spazio. Nel caso degli ambienti urbani, l'elevata variabilità spaziale dei fattori di rischio sanitario impone la necessità di ricorrere a un'elevata granularità spaziale, ovvero sia la raccolta dei dati che la loro analisi devono essere effettuate utilizzando metodi caratterizzati da una risoluzione spaziale molto elevata. Questo concetto è alla base di PULSE e dell'attività di ricerca presentata all'interno di questa tesi, la quale presenta un insieme di strumenti e ricerche svolte nel contesto di PULSE, grazie ai dati raccolti e al sistema sviluppato. Nel dettaglio, questa tesi presenta la maggior parte del lavoro svolto dall'Università di Pavia, manager della parte tecnica del progetto PULSE, con la collaborazione della New York Academy of Medicine e delle New York University. Dopo una breve introduzione e una descrizione del progetto nel suo complesso, vengono presentate alcune applicazioni di metodi innovativi ad alta risoluzione spaziale, concentrati principalmente su due argomenti: lo studio del panorama sanitario e i fattori di esposizione nell'ambiente urbano per la creazione di modelli di salute pubblica e la creazione di strumenti interattivi di simulazione per favorire lo sviluppo di strategie di intervento che possano migliorare la salute e il benessere generali. Nel corso del progetto infatti, alcuni metodi di machine learning abilitati spazialmente sono stati utilizzati per studiare l'andamento delle ospedalizzazioni per asma nella città di New York in relazione a un grande numero di fattori demografici, ambientali e socioeconomici, e i risultati di questo studio sono stati utilizzati per la creazione di modelli predittivi di simulazione, includendo al loro interno anche modelli di sistemi dinamici che rappresentano il traffico urbano. Inoltre, è stato condotto uno studio

basato sull'applicazione di metodi di deep learning in combinazione con algoritmi di clustering per raggruppare diverse aree urbane in base alla loro struttura caratterizzante e trovare le conseguenti relazioni tra struttura urbana e salute della popolazione. Un ulteriore studio è stato poi condotto a Pavia, in Italia, per determinare l'eventuale effetto del lockdown istituito per contenere l'epidemia di Covid-19 sull'inquinamento atmosferico della città, utilizzando una fitta rete di sensori sviluppata dall'università. I risultati di tutti questi studi evidenziano l'importanza dell'utilizzo di un'elevata risoluzione spaziale nello sviluppo di modelli descrittivi, predittivi e di simulazione nel contesto della salute pubblica nell'ambiente urbano.

Abstract (English)

According to the World Health Organization, about half of the world's population lives in big cities, and this proportion is expected to increase significantly in the nest decades. This demographic change, together with the fast technological progress, is causing an important variation in the set of environmental factors we are all exposed to in the course of our lives, and also a consequent change in our behaviors and lifestyle. Several studies have shown an increase in the prevalence of some respiratory and cardiovascular diseases in most of the urbanized world and hypothesized that the cause can be found in these changes. New fields of study are rising, such as *exposomics*, i.e. the study of the effects of all the external and internal factors we are exposed to and that can influence our health and wellbeing, that can be related to *public health*, i.e. the science of preventing health problems and improving quality of life through coordinated efforts by public and private organizations and single individuals.

In this contest, several public health projects have been created in the last years to exploit new technological advancements to improve the prevention of diseases. One example is the PULSE project, funded by the EU Commission in the year 2016 with the aim of creating a collaborative multi-technological system to prevent asthma, type 2 diabetes and cardiovascular diseases in big cities, using technology, big data analytics and geostatistics to assist both citizens directly and public health policy makers who can organize interventions on the urban territory. Big cities are very heterogeneous environments in which demographic, environmental and socioeconomic differences are often pronounced, and exposure factors can change significantly even within areas that are relatively close geographically. Several studies have demonstrated that social inequalities, high-pace lifestyle and air pollution conditions typical of urban environments have influenced the public health scenario causing an increase in

prevalence of certain diseases, including the ones treated in PULSE, more pronounced than in rural areas. For this reason, PULSE aims at providing a set of tools to increase awareness and ease the creation of prevention and intervention strategies to help mitigate health risk in big cities. To this end, the PULSE systems feature various technological components that connect both citizens and public health policy makers to the city's data. Particular importance is given to the spatial dimension, as public health data cannot be properly described and analyzed without taking into account the geographic component. In the case of urban environments, the high spatial variability of the risk factors requires to study the problematic with a high spatial granularity, i.e. both the collection of the data and the consequent analyses have to be performed with a high spatial resolution. This concept is at the basis of PULSE and of the research presented in this dissertation, that presents a set of tools and public health studies developed thanks to the data gathered in PULSE and its architecture. More in detail, this book presents most of the work done by the author of this thesis, that took part in several projects with his team at the University of Pavia, that has been the technical manager of the project, in collaboration with the New York Academy of Medicine and the New York University. After an introduction and a brief description of the whole project, some applications of high-spatial-resolution innovative methods are presented, focusing mainly on two paradigms: the study of public health phenomena and exposure factors in the urban environment for the creation of health models, and the creation of interactive and simulation tools to help designing interventions in the city. During the course of the project in fact, highly spatial enabled methods have been used to determine how asthma hospitalizations in New York City could be related to a high number of environmental and socioeconomic factors, and the findings of this study were used to create predictive simulation tools, including also traffic dynamics. Furthermore, deep learning methods were used in combination with clustering algorithms to group city areas according to their urban landscape and find relations with the health status of the inhabitants. An additional study was then carried on in Pavia, Italy, using a dense sensor network to determine whether and how much the Covid-19 lockdown had an impact on air quality in the different areas of the city. The results of all these studies, thoroughly described in this thesis, all

enlighten the importance of spatial enablement and high spatial resolution in urban public health modeling and simulation.

Contents

Chapter 1

Thesis Overview

Public Health is a novel field of study that generates from a multidisciplinary and complex concept, i.e. the idea that health, concepted as a combination of longevity, quality of life, prevention, social and psychological wellbeing, is the result of the combination of a large number of personal, social and environmental factors. Some define "Public Health" as the science of preventing diseases and improving quality of life through organized efforts that come from society, public and private organizations, individuals, politicians etc. [1]–[3].

This concept is strictly correlated to the concept of *Exposomics*, that can be defined as the science that studies the effect of all the internal and external factors every human being is exposed to during his/her entire life [4], [5]. As scientific, medical and technological progress advances, scientists have realized how human health is the result of a complex mixture of elements, that include environmental factors, genetic predispositions, personal choices of behavior, specific external factors. The combination of all these elements is able to affect our health status both with a psychological and physical impact.

As the 21st century society advances, due to the fast technological advancements, new exposure factors are constantly been created (e.g. new pollutants, E.M. waves, climatic changes etc.) and the socioeconomic equilibrium of the modern society is continuously evolving, resulting in new challenges in the public health contest. For example, several studies have demonstrated that in many parts of the world the prevalence of some multifactorial

diseases such as asthma, diabetes, genetic diseases and cardiovascular disorders has shown important trend modifications in the last years [6]–[8].

In this contest, new public health multidisciplinary projects are being funded and developed in order to study and solve these new problematics. The research work described in this thesis is based on one of these projects, named PULSE, that was funded by the European Commission and has developed new knowledge and a new set of tools to improve the public health status in the world's big urban environments, focusing on several topics such as air pollution, lifestyle, asthma, type 2 diabetes and cardiovascular diseases. The main purpose of this project was the creation of a collaborative multi-technological system to reduce and treat the risk of asthma, type 2 diabetes and cardiovascular diseases in the big cities, involving both citizens and public health authorities in a well-studied data exchange that integrates information coming from several sources.

This data exchange, that flows through several technological components that will be briefly described in this thesis, starts from the citizens, some of which are also patients of either one or more than one of the considered diseases, who can anonymously send their own data and geographic position and receive feedbacks and advice in return, that are generated by the big data and risk models integrated in the system. The other main recipients of information is represented by the public health authorities, that are connected to the system through the so-called Public Health Observatory (PHOs), through which they can visualize, analyze and inspect general data about the city and take informed decision with the aid of specific simulation and analysis tools.

Data integration is one of the main bases of this project and of the work presented in this thesis, that focuses on the creation of some specific innovative tools based on Big Data and geostatistical analytics to analyze aggregated data, predict public health outcomes and organize proper prevention and intervention strategies.

1.1. Big Data for Public Health

As already stated at the beginning of this chapter, Public Health is a highly multidisciplinary field, as human health is concepted as

the combination of a high number of factors. As a consequence, to study public health it is necessary to solve complex problems that involve the integration and analysis of enormous quantities of data characterized by high levels of heterogeneity, i.e. coming from different sources and diverse in typology, format, dimensional scale, temporal unit. Therefore, Big Data analytics are necessary to solve these kinds of problems.

The term *Big Data* has gradually become more common in the last decades, and yet it does not correspond to a universal definition. In spite of this, with this term, scientists usually indicate datasets that are so vast to make it impossible to visualize and analyze them with conventional methods, as they require dedicated algorithms and procedures instead [9]. Big Data are usually characterized using a set of terms that can be defined as *the 4 Vs* [10], as there are four different terms, all starting with a V, that are proper to describe them. These terms are:

- Volume – this term refers to the high quantity of data that must be analyzed and processed. Some datasets can contain millions or billions of observations.
- Velocity – thanks to the new technological advancements in this field, large quantities of data are collected in short time, often within seconds.
- Veracity – in order for them to be useful, big data have to be representative of a real phenomenon, so it is important to avoid or correct errors of various kind as much as possible during the whole process, from the collection to the analysis.
- Variety – this word refers to the implicit heterogeneity of these data, since they usually come from different sources and represent different phenomena.

When it comes to studying the human body or the human health, all these concepts are particularly emphasized, as the human body itself is a complex system that interacts with the external world in unforeseeable ways and provides enormous quantities of data characterized by high variety. Just to give an example, genetics alone is able to provide high quantities of data (gene expression of more than 25,000 genes, proteomics, metabolic pathways), that usually have to be integrated with clinical data, environmental

factors, data characterized with a highly variable dimensional scale, that goes from the intracellular to the population scale.

Public Health makes no exception under this point of view, since community health is the result of different levels of health, starting from the single person, and strictly depends on the *Exposome*, i.e. the combination of all the internal and external factors every person is exposed to in a lifetime.

1.2. Outline of the Thesis

This thesis reports a portion of the work performed during the PULSE project by the University of Pavia, presenting several tools and research projects that show how some of the most innovative public health and Exposomics concepts and tools have been developed inside the project. The specific role of the PhD candidate authoring the work has been to take part both to the experimental design and the analysis part of all the studies presented in this work, specifically the ones regarding asthma risk assessment using a combination of environmental and sociodemographic factors and the creation of simulation tools, which have been almost entirely designed by the author. The other projects reported in the thesis have been carried out by the technical team of the University of Pavia, with the active participation of the PhD candidate both in the design and the development of the methodological parts, and also in the evaluation and interpretation of the results.

The next chapter, **chapter 2**, presents the research background of the project, focusing on the motivations that lead to the creation of the project itself and of the tools that are presented in the rest of this dissertation. Some related projects are presented as well.

Chapter 3 presents the PULSE project in a brief but complete way, showing how it was created and what is the mail architecture and data flow from a practical point of view. Some external resources that have been useful for the work presented in the following chapters are introduced as well.

Chapter 4 and **chapter 5** are both methodological chapters, that present the main methods and algorithms that have been used in the projects that are described in this work. In particular, chapter 4 focuses on the spatial analytics that have been fundamental in the creation of spatially enabled algorithms that allow to perform

proper visualization, analysis and prediction on health in the urban environments with a high spatial resolution, whereas chapter 5 explains the technology used to perform interactive simulations to explore possible public health scenarios, reusing the spatial analytics previously introduced.

Following the same idea, the main projects and applications of the proposed methodologies are presented in two chapters, i.e. **chapter 6** and **chapter 7**. The first one presents the main applications of the spatial analytics introduced in chapter 4, showing the results of three different studies where spatial enablement has been of fundamental importance to create innovative ways to address urban public health; the second one shows some example of applications of the simulation tools presented in chapter 5, showing how they can be used to perform predictions on the health status of the citizens integrating environmental, demographic and socioeconomic data.

Finally, **chapter 8** contains a few conclusions and possible future developments in the research field this dissertation is inserted in.

Chapter 2

Background and Objectives

In the first chapter of this thesis, important concepts such as Public Health, Exposomics, Big Data have been introduced and their connections explained. In particular, it has been explained how, in most cases, studying Public Health problems cannot be properly done without the use of Big Data methods.

In this section, the scientific and social background of the research reported in this manuscript will be better explained, pointing out important elements that lead to the idea of the PULSE project and in particular to the necessity to focus on the methodologies and studies reported.

2.1. Health in the Big Cities

According to the WHO, about 54% of the world's population is currently living in urban environment, and the tendency to concentrate in big cities is expected to continue, as in 2030 big cities will probably contain the 60% of the global population and this percentage is projected to increase up to 66% in 2050 [11]. Big urban sites are notoriously heterogeneous environments, where social, economic, demographic and environmental contrasts are particularly emphasized, since large differences in all these fields can occur in relatively small spaces. For this reason, big cities are one of the most interesting sets for Public Health studies, as citizens are constantly exposed to a dynamic environment, and external factors such as air pollution and lifestyle can have an important effect on their health and quality of life. Plus, some

factors that have been demonstrated to have an important influence on human health and life quality are peculiar of big cities, such as noise, garbage collection, safety etc.

In this context, several recent studies have pointed out a change in trend of the prevalence of several respiratory and cardiovascular diseases in the industrialized world, particularly asthma [12], pulmonary infections and allergic rhinitis [13]. These disorders appear to be becoming more common in several areas of the industrialized world, and this could be partially a consequence of the change in exposure factors that the new lifestyle created by the expansion of the urban environments. This is another reason why urban public health is gradually becoming an important field of study for the scientific community, and public health study projects on this topic are being created and developed. PULSE, introduced in Chapter 1 and better explained in Chapter 3, is one of these projects, and focuses mainly on two problems: the increase of asthma related to air pollution and the increase of type 2 diabetes and cardiovascular diseases related to lifestyle. These diseases, although widely known and studied by the medical community, are usually treatable but not always curable, therefore prevention is the best strategy to contain their effects. To perform a proper prevention strategy, it is important to understand in detail which elements are decisive in causing the diseases and how do they work. This can be done only collecting and analyzing a large amount of data to discover the disease patterns in the different areas of the city. For this reason, studies that address the mechanisms that are leading to an increasing prevalence of asthma, type 2 diabetes and cardiovascular diseases in the urban environments are not common, as performing data collection at an intra-city level is not always easy and could lead to a lack of a sufficient quantity of data to obtain significant results. Therefore, a proper prevention strategy can be designed only after the definition of an effective data collection paradigm.

Another important aspect of urban public health is cooperation between individuals, as a big city is a heterogeneous environment where a lot of people live in a small area, and the lifestyle and choices of one person can influence the ones of the others, as every individual contributes to creating a peculiar environment which every inhabitant is exposed to. For this reason, information and communication are other important elements in the creation of a proper prevention strategy. The ideas of PULSE can be inserted in

this context, as the project aims at creating a paradigm that allows to improve data collection and analysis and simultaneously promotes information and cooperation between entities.

In the next sections, the main problematics at the basis of PULSE are explained.

2.1.1. Air Pollution and Asthma

Asthma is a common condition in which human airways swell and sometimes tend to produce extra mucus, occasionally leading to difficulties in breathing, cough, and other respiratory disturbances [14]. In most cases it is not lethal and it can be considered only a minor nuisance, but sometimes it could be a more serious problem that interferes with daily activities, and in some people it can have severe complications with attacks that require immediate medical assistance and even hospitalization, as the condition can become life-threatening [15]. There is not a definitive cure for asthma, but symptoms can be treated, and in some cases severe attacks can be prevented. Asthma attacks in fact can be triggered by specific factors such as dust, cold air, chemical substances, pollens and pollution particles [16], so if the trigger is known the worsening of the condition can be prevented by reducing exposure to it.

Being a very common disease, even though generally not severe, asthma represents a considerable source of costs and consumption of resources for the public health systems, and a lot of recent studies and projects are studying it in order to reduce the prevalence and find new cures. Several studies have demonstrated that the prevalence of asthma has been increasing in the last decades in several parts of the world [6], and in some areas it represents a huge problem.

The existence of a link between asthma and outdoor air pollution has been widely demonstrated [17], as a consequence, an increase of air pollution in the urban areas is expected to lead to an increase of asthma-related problems in the population. Almost all the pollutants have shown a link with asthma, particularly sulfur dioxide [18] and particulate matter (PM2.5 and PM10) [19]. Pollution can trigger asthma with two different mechanisms, as it can cause short-term and log-term effects. Short-term effects are mainly referred to asthma attacks and complication related to

immediate exposure to high levels of pollution, that cause irritation of the airways of susceptible individuals leading to an asthma outcome; long-term effects are those due to prolonged exposure to pollutants (that can last days, months or even years), that causes a progressive deterioration of the airways status that increases the probability of asthma attacks and reduces the easiness of recovery from asthma complications.

Although all the mechanisms leading to these processes are not entirely known, it is clear that in order to reduce health risk related to asthma in cities it is necessary to reduce the presence of asthma triggering factors such as air pollution. Several steps have been taken in this direction in the last decades throughout the world (traffic limitations, laws that regulate emissions of factories, ban of diesel cars etc.), but the generalized lack of an organically unified response leads to limited benefits and highlights the necessity to further intervene. The PULSE project, described in this thesis, proposes a new paradigm of intervention based on the direct involvement of the population.

2.1.2. Measuring Air Pollution

Planning a proper intervention on air pollution is a process that starts with an accurate measurement of the concentration of pollutants. In order to gain a sufficient quantity of air quality data to perform an accurate study on the effects of pollution on personal and public health, measurements have to be taken with a high spatial and temporal granularity, and errors should be reduced to minimum.

As a consequence, the quality of the sensors or monitoring stations used is crucial to obtain proper measurements. Although the new sensing technologies allow to easily have high quality sensors and share data with high temporal resolution, some issues depending on the ratio between cost and accuracy of the sensors are still unresolved. As pointed out by some recent studies, the main problem in measuring air pollution in urban environments comes from the usually relatively small number of monitoring stations that are deployed on the territory, that is a consequence of their high costs and demands in maintenance. For example, a large city like New York, that is developed on an area with 784 km^2 of surface has only 13 official monitoring stations, and not all of them

measure the same pollutants; the city of Pavia, extended on a 62.86 km^2 territory, has only 2.

A small number of monitoring stations can lead to dangerously neglecting local variations of the pollution concentration, that in urban environments can be very pronounced, as the conditions concerning traffic, emissions, factories etc. can vary significantly even in a small environment. To overcome this problem, low-cost sensors are becoming more common every day [20], [21]: these sensors are usually small, portable and much less expensive if compared with the official monitoring stations, they also require less maintenance, but the accuracy of their measurements can be significantly lower than the one of the more expensive stations, and they usually require a proper calibration process.

Considering all these facts, the best strategy to obtain the best measurements in terms of quality and spatial granularity is to find a proper trade-off combining the use of high quality monitoring stations and low-cost sensors, in a way that maximizes accuracy and spatial resolution together. Numerous projects have been created to move in this direction [22], [23], and this topic is analyzed and applied also by PULSE (see next sections), as reported in this dissertation.

2.1.3. Cardiovascular Diseases

The term *cardiovascular diseases (CVDs)* indicates a class of diseases of different type that concern the heart and/or the blood vessels [24]. This definition includes congenital conditions as well as degenerative ones, some examples are coronary heart disease, cerebrovascular disease, vein thrombosis, congenital malformations leading to risk of heart failure or vascular problems. According to the WHO, cardiovascular diseases are the first cause of death globally, as people die from them more than from any other cause. For example, it has been estimated that in the year 2016 about 17.9 million people died from CVDs, representing the 31% of all global deaths of the year. Among them, 85% were due to heart attacks or strokes. The prevalence of these conditions is particularly high in the low- and middle-income countries, although high-income areas also present a significative burden in hospitalizations and deaths due to these diseases.

CVDs are widely studied diseases and there is a general consensus on the fact that most of them can be prevented acting on risk factors that can be both environmental and behavioral, such as tobacco use, obesity, lack of physical activity. Some people are more at risk of cardiovascular complications than others, especially those with genetic predispositions or pre-existent diseases like hypertension or diabetes, most of which can be prevented themselves reducing exposure to risk factors and with behavioral changes.

Recent research has shown a tendency of several risk factors to increase in the last years in several parts of the world, for example obesity and hypertension [25]. This increase can be imputed to a general change in lifestyle that occurred in the fast-pace society we live in, that is particularly noticeable in big cities, were working hours and stress often lead to incorrect nutrition patterns and/or inconstant physical activity. Plus, in some cases information about this topic is not proper diffused among the population. Several studies showed that the prevalence of heart and vascular problems is higher in low-income areas, and there is a correlation between limited economic possibilities and obesity [26], that is probably a consequence of the higher cost of healthy food and the lower availability of healthy stores or restaurants in the low-income areas. For this reason, the reduction of cardiovascular diseases risk must be a combination of personal behavioral change and community urban planning that can help the citizens to easily access services that can help them stay healthy.

2.1.4. Type 2 Diabetes

The term *diabetes*, rather than a specific pathology, indicates a class of conditions that lead to an abnormally high blood glucose level [27] that can be dangerous for several organs and biological functions. In fact, dangerously high glucose concentration peaks can lead to a condition called ketoacidosis that can lead to coma and be life-threatening, and a prolonged situation of high blood glucose can lead to severe complications such as exposure to infections, coagulation problems and blood vessels damages that can permanently affect some organs such as the limbs and the eyes, causing even vision loss or necessity to perform limbs amputations.

There are mainly three types of diabetes known to the medical community:

- Type 1 Diabetes: this condition is an autoimmune disease [28, p. 1] that occurs when the immune system, for causes that are still mostly unknown, attacks and destroys the β-cells located in the pancreas, that are responsible for the production of insulin, a hormone that regulates the glucose level in the blood. This condition typically occurs in young patients, although it could also hit adults and elderly people, and it causes a total lack of insulin, therefore the only way the patient can lower his/her glucose levels is by injecting insulin inside his/her blood stream. This type of diabetes is much less common than diabetes type 2, as according to the ADA (American Diabetes Association) [29] about 5% of all diabetes diagnoses are type 1. The causes of type 1 diabetes are not entirely known, it has been recognized that there could be a combination of genetic predisposition (the disease appears to run in families) and unknown environmental factors.
- Type 2 Diabetes (T2D), differently from type 1 diabetes, is not an autoimmune disease. T2D happens when, due to a combination of genetic predisposition and specific risk factors [30], the body develops a resistance to insulin, so the cells cannot exploit it properly even if the pancreas is able to produce it in a sufficient quantity. This condition is usually preceded by a transition phase named *prediabetes*, that defines a condition in which the blood glucose levels are not yet dangerously high, but there are signs of altered concentration control such as peaks slightly higher than normal or a basal level that tends to be high [31]. When the glucose levels are slightly but not extremely high, such as in prediabetes or in an initial stage of T2D, the patient does not experience any symptoms, creating a dangerous situation as the damaging effects to the organs and blood vessels can already start in this phase. Therefore, in order to limit complications, prevention and screening of T2D are very important.

T2D cannot be cured, even though it can be treated by some specific drugs or with extra insulin intakes (in some cases), so prevention and risk factors limitation are the best way to act against it. T2D tends to be more common in adult patients, with its prevalence increasing with age after 45 years.

- Gestational Diabetes is a particular kind of diabetes that happens specifically in women during pregnancy. As in T2D, this condition causes an increase of blood sugar levels due to a different sensitivity of the body cells, but the condition is generally temporary and tends to fade after delivery [32]. Not all the causes of this condition are known, but scientists assume that the leading cause is a change of the hormonal equilibrium in the pregnant woman's body that somehow interferes with the normal glucose metabolism. Risk factors for gestational diabetes are mostly similar to the T2D ones and include obesity, lack of physical activity, smoking, advanced age etc., and taking action on these risk factors can prevent gestational diabetes to worsen.

Diabetes is probably one of the most studied diseases of the modern times, as it is widely diffused and its complications are responsible for widespread conditions that often lead to hospitalizations, reduced quality of life and increased mortality. Several studies have pointed out that the prevalence of T1D and T2D is increasing in most of the developed countries. The reasons of this are not entirely understood, but in the case of T2D lifestyle changes could play a role in the same way it happens for cardiovascular diseases and asthma. This process also concerns the big cities, in fact past research has pointed out that two thirds of the people having T2D live in big cities [33], and this is thought to be the consequence of mainly two factors: with the new century prosperity, people tend to live longer than the past, and with the lifestyle changes brought by the modern society several risk factors such as obesity are increasing, especially in big urban environments [34]. Several city councils and local organizations are taking actions to face the spread of T2D, with campaigns that aim at informing the population or intervene on the urban environment in order to encourage physical activity.

As it happens for asthma and CVDs, T2D is more common in low-income areas, and the probability to have severe T2D complications that lead to hospitalizations and irreversible damages such as vision loss or amputations is higher in the same areas [35]. This is probably related to the fact that risk factors such as obesity and smoking often present positive correlations with poverty. Therefore, also in this case some urban areas are more in need of interventions than others.

2.2. Related Projects

The increasing awareness of the public health problematics typical of the urban environments, leading to new challenges in this field, has led to the creation of a new set of public health projects focused on health and wellbeing of the population in the big cities. The PULSE project, which all the work reported in this thesis is based on and that is presented in the next chapter, is part of a large cluster of projects that constitute a European framework named Horizon 2020. This framework includes numerous projects aimed at performing technological innovation and research on a lot of modern topics such as public health, global warming, food safety, water supply, elderly population etc.

Many of these projects focus on the topic of health and wellbeing in the urban areas. For example, the project *City4Age* [36] aims at facing the new challenges deriving from the ageing population creating a hospitable urban environment for the elderly, when they can have assistance through new technologies through which they can detect early risks related to frailty and/or mild cognitive impairment and receive personalized interventions to improve their quality of life and be encouraged to maintain positive behaviors. This project works in close contacts with city councils and communities to facilitate the roles of social and health services. This project pairs with other European projects, such as *GRAGE* [37], that aims at creating a better social environment for the ageing population in the big cities, fostering innovation in themes such as green buildings, food delivery, technology, information and language.

Another example of urban public health project aiming at facing the new challenges related to demographic changes is *Urban GreenUP* [38], that aims at developing a methodology to support

the development of renatured cities in the context of mitigating global warming and improving the quality of life of the population, as many studies suggest that a city without green areas is related to a higher prevalence of depression, mental diseases and diseases related to air pollution.

The topic of urban air pollution is treated by several projects as well, for example *ICARUS* [39] aims at developing innovative tools to assess the impact of climate and demographic changes on urban air quality and to support new policies to contrast the air pollution increase and the consequent damages on the population, taking into account also socioeconomic factors and analyzing population subgroups. Ina similar way, *iSCAPE* [40] aims at integrating and advancing the control of air pollution in the urban environment through the development of sustainable emission reduction strategies, policy interventions and behavioral change initiatives.

These projects are just a few of hundreds of examples that could be made of projects treating the difficulties deriving from the changing times and the increase of population in the urban areas. All these research initiatives are based on performing innovation through technology and Big Data analytics, building solid infrastructures based on information, policy making and collaborations with the communities.

PULSE is entirely inserted in this context, as it was born with the same ideas and to face problematics of the same kind, but focusing on slightly different aspects, i.e. the prevention of asthma, type 2 diabetes and CVDs. This project is described in detail in chapter 3.

2.3. Limitations in Public Health Studies and Objectives of the Thesis

Although the awareness on the necessity of performing novel public health studies related to urban environments has been increasing in the last years, even the most recent research suffers from several limitations, mainly related to the lack of data to perform proper analyses and interventions at a sufficiently high spatial resolution. As it will be shown in the methodological section of this thesis, the heterogeneity of the environments and the population's characteristics inside big urbanized areas makes it

crucial to study public health phenomena at a neighborhood level, as the exposure factors can change drastically from one area to the other, even within limited geographical distances. Due to a general lack of data collected at a high level of spatial granularity, studies that address urban public health at a sufficient level of geographical detail are still rare, as for the most part the whole city or extended areas of it are condensed together, possibly hiding important local phenomena. For this reason, one of the main concepts at the basis of PULSE, and above all at the basis of the research part of PULSE that has been carried out by the author of this thesis, is the importance of considering the spatial dimension in the definition of visualization, analysis and intervention tools for urban public health. Through the application of highly spatially-enabled methods it is possible to assess all these phases of public health problem solving with a high spatial granularity, taking into account also local situations that could be neglected with other experimental settings. Of course, to make this possible it is necessary to possess data collected at a high spatial resolution, which still represents the weak point of urban public health. Steps forward in this direction are being made in some local realities (e.g. in New York City, as it will be shown in the next chapters) and thanks to specific projects such as PULSE, that provides also innovative data collection methods that aim at solving this problem. Therefore, the work described in this thesis shows the importance and the opportunities given by the inclusion of spatial enablement in the design of urban public health studies, both on the analysis side and on the intervention side, providing novel ideas on how to face the new challenges raising in urban health working at a neighborhood level.

Chapter 3

The PULSE Project

This chapter is dedicated to a detailed presentation of the PULSE project, its principles, its rationale and the main technological components.

PULSE, an acronym that stands for Participatory Urban Living for Sustainable Environments, was an international project funded by the European Commission under the Horizon 2020 framework [41], it started in the year 2016 and ended in 2020. It involved several partners, both in the academic world and in the private sector, from all the world. Each partner participated with their distinguished expertise, forming a highly multidisciplinary environment that allowed to create a complex system based on the integration of heterogeneous datasets and technologies.

In details, the partners that took part at the project are:

- Universidad Politécnica de Madrid (UPM): this institute is the largest technological university of Spain. Among the research groups, they have a large team specialized in consultancy, design, development and deployment of eHealth solutions including reliable and effective telehealth services, personal systems for self-management of health and integrated regional

information systems. UPM has been the coordinator of the whole project and has been responsible for the development of the Pulsair App.

- Università di Pavia (UNIPV): the University of Pavia is an important Italian institute located in northern Italy. Inside UNIPV there are a large center of bioengineering that includes a data analysis core, and a geomatics lab specialized in georeferenced data, geostatistics and spatial enablement. UNIPV contributed to the project with two different teams, one coming from the geomatics lab that has been in charge of integrating data in the WebGIS and of spatially enabled analyses, the other formed in the biomedical informatics lab that has been in charge of data analysis and integration for the dashboards and has been the coordinator of the technical team of the whole project.
- Università di Padova (UNIPD): the university of Padova is another historical Italian institution. Its department of Information Engineering contains a bioengineering center that has one of leading groups in the field of mathematical modeling of diabetes and its complications. UNIPD has been the leader of the risk modeling part of the project.
- European Connected Health Alliance (ECHA): ECHA is the trusted connector, facilitating multi-stakeholder connections around ecosystems, driving sustainable change and disruption in the delivery of health and social care. The main task of this group has been facilitating the dissemination, exploitation of results and communication. ECHA has been also in charge of organizing meetings and events, attracting external partners and encouraging collaborations.
- The New York Academy of Medicine (NYAM): this historical center of research, located in the city of New York, is specialized in projects and research that favor health and wellbeing in the urban environments of NYC and the world. NYAM contributed with its Center of Health Innovation coordinating the activities of the test site in New York, collecting and providing data, statistics and analyses useful for the system, and

cooperating with the technical team to develop risk models of asthma and diabetes with specific tools.

- Belit Ltd.: this private company, founded in Belgrade, Serbia, is specialized in software development, C# and Java coding for enterprise applications, database development and treatment, data processing and management. Belit has been in charge of the development of the database structure and the raw data management, plus the development of the dashboard by integration of data and external tools developed by other partners.
- Public Health Agency of Barcelona (ASBP): ASBP is the main public health provider of the city of Barcelona, co-participated by the Barcelona city council and in cooperation with the Catalan government. The main aim of ASBP is monitoring health and wellbeing in the city, designing and implementing intervention strategies and policies to improve them. ASBP has been responsible of the definition of public health indicators and has cooperated with the definition of diabetes risk models. It has also coordinated the Barcelona test site.
- GENEGIS: this is an Italian company, with headquarters in Milan, that is specialized in Geographical Information Systems (GIS), and more precisely in the creation of solutions for the integration of geographic information in information systems. In PULSE, GENEGIS had the main task of implementing the spatial information framework and creating and maintaining the WebGIS.
- Birmingham City Council (BCC): BCC is a large local authority in charge of creating policies to support the vast population of the metropolitan area of Birmingham, UK. It includes a new framework named *Digital Birmingham*, that contains policies to invest in the use of the new technologies and digitalization to ensure benefits to the whole population. In PULSE, BCC has been the leader of the Birmingham test site and has contributed to the creation of the Public Health Observatories.

- Institut Mines-Télécom (IMT): IMT is a public institute under the authority of the French Ministry of Industry and Electronic Communication, focused on education, research and innovation in engineering and digital technology. The main headquarters are in the metropolitan area of Paris, France, but IMT has many associated laboratories in other areas of the world, one of which is located in Singapore. In the project, IMT has been the coordinator of all the test beds and has been directly in charge of the Singapore one. It has also contributed to the integration of public health models and maps thanks to its IT expertise.

3.1. Main Concept and Purposes

As already stated in the introduction section, PULSE is a Public Health project with the main aim of the creation of a collaborative system to prevent and treat asthma, type 2 diabetes and cardiovascular diseases in the big cities. This aim is achieved through the use of new paradigms of technology exploitation, based on the direct participation of the users who the benefits of the project are intended for.

The main pillar of PULSE is data integration, as different systems with different purposes and based on different scientific backgrounds work together in an advanced data exchange, and create an advanced instrument that allows both users and public health authorities to cooperate in improving the community health encouraging behavioral change and urban planning.

PULSE can also be defined an Exposomics project, as health risk is understood to be a complex combination of personal, environmental and socioeconomic factors, with an added role played by human behavior. For this reason, an international equipe of experts in different fields (engineering, informatics, sociology, marketing, psychology etc.) has been reunited to create the PULSE concept.

Among all the elements that concerned the studied diseases, there were two main clinical focuses in the project: the link between air pollution and asthma [17] and the link between physical activity and type 2 diabetes [42]. Although the existence of these links is common knowledge nowadays, the increase of

prevalence and complications in these two diseases showed the necessity to define new intervention strategies to allow for the urban areas' population to assume the right behavior to improve the community quality of life through information campaigns, assistance in the every-day life and dedicated urban planning strategies from the proper authorities. All these elements were considered and developed in the different parts of the project.

Another element that has been pivotal during the definition and development of the project has been the importance given to spatial dimension, and in particular to the necessity to use a high spatial granularity in the definition of the urban public health problems and the relative solutions. One of the main gaps of the scientific research conducted on urban health in the recent years that was found during the project was the general lack of studies that address urban health at a spatial resolution sufficient to properly spot all the peculiarities of the cities at a local level. One of the main characteristics of big cities is the high variability of environments and populations that reside in a small space, thus studying health considering the whole city as one environment or using large spatial subdivisions can be risky, as it can result in hiding some important local variations of health or quality of life conditions. The problem with this issue is that in most cities data are not gathered in a sufficient quantity to perform meaningful statistical analysis. Among other things, PULSE proposed a model of technological environment that allows to overcome this issue, with interventions at different levels: as personalized feedbacks help the users singularly, the users provide useful information that can be used to better analyze the communitarian health status of the city at a neighborhood level, thus providing the health authorities with more powerful tools to understand the city's problematics and design the proper intervention strategies.

After the definition of the project and the technological features, the system has been tested in seven cities in three continents: Barcelona, Birmingham, Keelung, New York, Paris, Pavia, Singapore. Each test site was chosen according to some peculiar characteristics interesting for the project such as the prevalence of certain diseases (e.g. asthma is very common in New York [43] and diabetes is a diffuse clinical problem in Singapore [44]) or known environmental problems (e.g. air pollution is a real problem in Pavia as in all the Po Valley [45]). In each one of these cities, a

large number of users has been recruited in order to proceed with the experimentation.

In the following sections, a detailed description of the technological architecture and the main components of the project is given, showing how different kinds of data are integrated in order to create a comprehensive system that includes useful tools both for the users and the public health policy makers.

3.2. Architecture and Big Data Infrastructure

The main elements of the PULSE architecture are: a personal App for the users, an innovative WebGIS, a set of dashboards for the public health authorities and back-end systems to connect all these elements.

The interface between the system and the users is represented by the personal user smartphone App, that is both a data collection tool through which citizens' data are gathered to be analyzed, and a receiver for the users that allow them to be connected to a set of useful tools and to receive personal notifications and risk scores. Another important tool is the WebGIS, that concentrates all the most advanced geographical tools and analytics used in the project in a powerful visualization tool. On the other end of the system, a set of dashboards allow the Public Health authorities to inspect the health situation in the city and some active interactive tools support them in the decision making process necessary to plan interventions and urban modifications.

It is important to notice that the App is not the only data collection tool in the project, as physical activity data is collected from the users through a FitBit device and a lot of external sources are used to gather health or socioeconomic data useful both for visualization and for the analysis used to provide feedbacks and information to the users and the policy makers.

Figure 3.1 represents the structure of the architecture, highlighting the data flow and the feedback mechanisms. The data flow inside the system is complex and includes a number of external elements such as air quality sensors, satellite images, open data sources.

More in detail, the data flow starts from the user themselves through the App. Some importance to the geographic dimension is already given at the very beginning, i.e. when the user signs up to his/her PULSE account, as the zip code of residence is one of the requested fields to fill out. Thanks to this geolocation paradigm, the user will be able to visualize maps, air quality records and receive personalized feedbacks that take into account the geographic environment which the user is immersed into. The user can also willingly consent to be tracked through the GPS tracking functionality in order to access some extra functionalities, such as the personal exposure calculator, that estimates the total air pollution intake considering the user's movements across the city and the interpolated air quality data (see section 7.3).

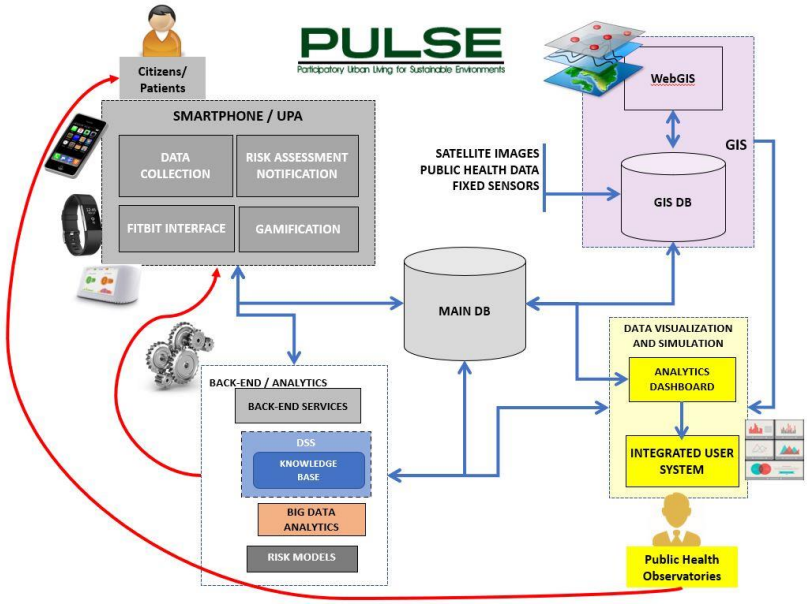


Figure 3.1: representation of the architecture of the PULSE system with the data flow between the components. The red arrows indicate the feedback mechanisms moving from the back-end to the App and from the Public Health authorities to the users.

The most obviously spatially enabled system in the architecture is the WebGIS, as it represents layers of data of different kinds all characterized by a specific spatial description. The data represented

is all stored in a specific area of the PULSE database into tables and structures where every entry is characterized by an element that indicates its spatial position in the real world.

The same spatial reference is maintained in the dashboards, where population statistics, maps and satellite images are all shown with a spatial reference and interactive tools have contain a spatial description or even spatial analytics, as illustrated in the next sections. Through the back-end services, the data flow continues from the App to the dashboards, generating two different feedback mechanisms, visible in figure 3.1: the first one starts from the risk models and calculators contained in the back-end services and gets to the user carrying risk scores and advice, the second goes from the dashboards to the user, representing the intervention and communication strategies that the public health policy makers can design in order to improve health and wellbeing in the city. All these feedback tracks are highly spatially enabled, as risk models take in input also variables related to the location the user is in (e.g. air pollution), and interventions by the authorities can be organized according to the results of spatial analytics that show where are the criticalities that need them.

The external data sources are for the most spatially enabled as well, since sensors data come from sensors whose position is georeferenced, satellite images are georeferenced as well and open data often contain elements that characterize the spatial dimension in which they were collected (e.g. zip code of patients whose census data is taken, the address of the hospitals where hospitalizations are recorded etc.). The next subsections describe the most important architecture elements more in detail.

3.2.1. User Personal App

The User Personal App (UPA), named Pulsair, is an App for smartphone created for the recruited users and it serves as their interface with the internal part of the projects. The UPA is a data collection tool, an information center and an active intervention tool at the same time. The recruited user can access the app subscribing with a personal account that can be created using an access code given by the PULSE administrators. After registration, the user signs an informed consent form where he/she is informed

about data collection and treatment, privacy rules, purposes of the project etc. All the collected data are anonymized and nobody beside the user can see their own data and recognize the individual who they were collected from, in compliance with the GDPR European data treatment rules.

After the login, the user is provided with a set of useful tools that interface him/her with the system allowing data collection, feedback reception and gathering of information. The main utilities of this system are:

- Health and wellbeing questionnaires: in order to gather all the data necessary for the calculation of personal health risk and the creation of personalized feedbacks, a set of questionnaires (13 including the general user information) has been created and integrated in the app. These questionnaires concern several topics, such as general information (age, sex, basic characteristics etc.), specific risk factors for one of the diseases, environmental contest of the user's place of residence, wellbeing indicators that include happiness, sense of direction, precepted quality of life. All the questionnaires are validated and widely used in literature for the computation of health indicators or risk scores.
  In order not to excessively increase the burden on the users, a notification with an invitation to complete one of the questionnaires is sent every few days, and the user is rewarded with an advancement of category and with useful information.
- FitBit interface: the personal App is not the only way data are gathered from the users, as physical activity indicators and some health data related to behavior and life habits (sleep hours, heartbeat etc.) are collected through an interface with a set of FitBit devices, which some users are provided with. In order to facilitate data visualization also for the users themselves, Pulsair features a page with statistics and summary of the data gathered during the last week of activity by the FitBit.
- Information about the city: when the user creates the account, he/she is asked to state which is the pilot site of residence. In this way, some useful information

about the city can be visualized, specifically information about air quality and meteorological conditions. Informative maps of the city can be visualized as well.

- Information about the project: useful information about the project, the aim, the consortium and the scientific rationale is given upon registration, but is also accessible in a dedicated section of the app, so that each user can always be well informed about the project he/she is taking part to.
- Feedbacks and advice: as already stated, the app is not only a data collection tool, but also an interface for the users with the system through which they can receive useful feedbacks and advice to improve their quality of life with the right behaviors. After the data have been gathered with the questionnaires, a personalized health risk score for each one of the diseases is calculated and shown in dedicated page of the app, where the user is informed about the level of risk (high, medium, low) of developing or worsening asthma, type 2 diabetes and cardiovascular diseases. Each level of risk is associated with a feedback message containing useful advice regarding the behavior to follow in order to lower the risk. Health risk is periodically recalculated and feedbacks are consequently retuned according to the new data gathered with the FitBits and with the repetition of the questionnaires.
- Gamification: in order to help the users to keep engaged with the App, besides the rewards coming from the useful advice and the consequent health improvement, a gamification paradigm has been applied, introducing levels of expertise inside the system in a way that awards the most active users. A rank of the users with the highest levels can be visualized in order to introduce a challenge effect that has been demonstrated to make users more willing to increase their engagement with the App.

With these systems, the UPA is a simple and intuitive system that serves two main purposes: data collection and advice deliverance. The layout of the App is also to studied to be pleasant

and easy to use. Figure 3.2 and 3.3 contain some screenshots of the app, showing the city information page and the health feedbacks page respectively.

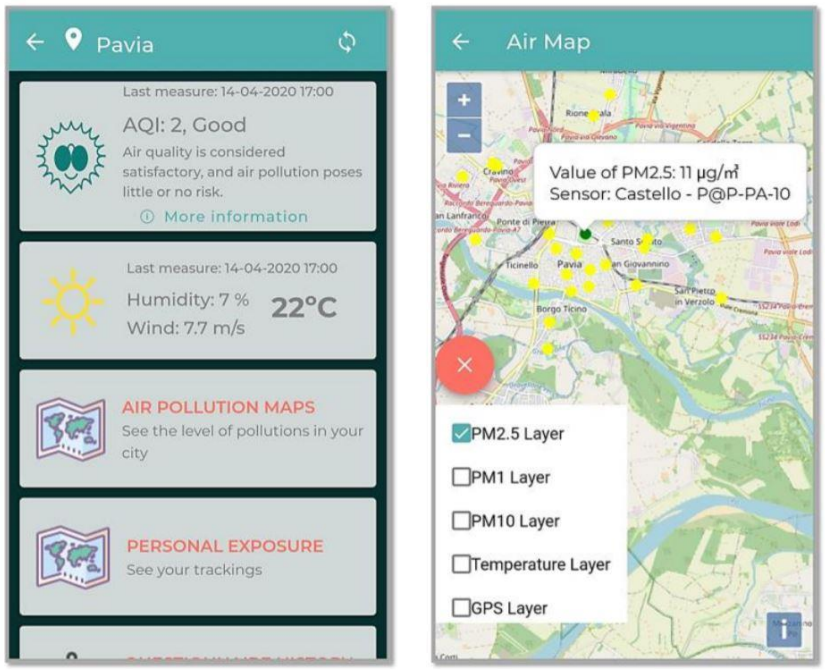


Figure 3.2: example of the city information that can be visualized inside the App. Air quality and weather data can be visualized on a general view with some associated feedbacks or on a map.

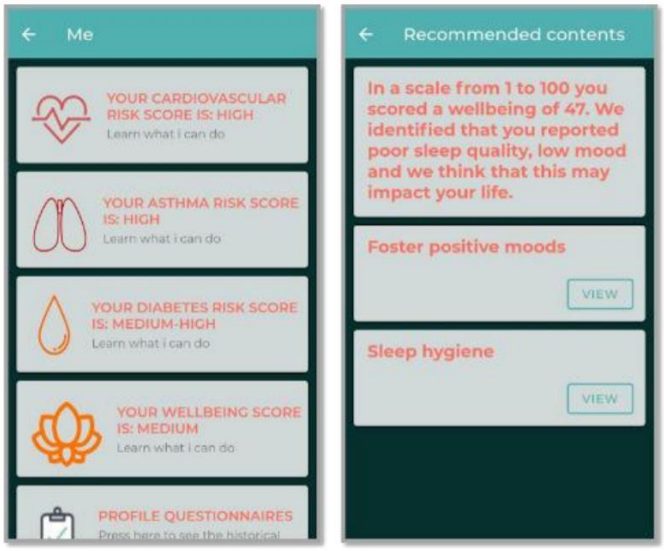


Figure 3.3: Examples of health-related feedbacks on the App.

3.2.2. The PULSE WebGIS

The word GIS (Geographic Information System) indicates a system designed to store geographic information and tabular data together, in order to allow the users to visualize data of different kinds with their geographic description [46]. A WebGIS is basically a GIS that is accessible online through a web browser. All GIS present typical features: the data is usually organized in layers, each one representing the information contained in a specific data table with its geographic description (usually with maps); layers can be switched on and off, overlapped and made transparent in order to visualize more phenomena together and how they change throughout a specific geographic area, facilitating pattern discovery or data analyses.

One of the main features of PULSE is the collection of a large quantity of data regarding the pilot sites, not only from the citizens, but also from external sources such as public health repositories, satellite images, air quality sensors, users' GPS tracking. All these data have a clear spatial reference, as they are collected in several spots throughout the cities' boundary and can show differences

over the territory crucial to proper analyze public health criticalities and risk factors distributions.

Plus, as already mentioned, one of the main paradigms at the basis of PULSE is spatial enablement, i.e. the ability to add a spatial dimension to the data gathered in a specific contest. Public health data are usually implicitly spatially enabled, as data regarding prevalence of diseases, distribution of risk factors, census of population, socioeconomic factors, air pollution etc. are usually collected from specific sources located in a recognizable environment.

For thee reasons, the PULSE WebGIS can be considered one of the main features of the PULSE system, as it also presents many innovative features. As in the whole project, data integration is one of the fundamentals of the WebGIS as well. The data collected in the seven pilot sites that can be visualized in the WebGIS concern mainly the following categories:

- Air Quality: air pollution maps are one of the most important features of the PULSE WebGIS. Air pollution data are usually gathered from fixed air quality monitoring stations, some of which were acquired by the consortium during the project and used in combination with the high-quality official monitoring stations to increase the spatial resolution of the measurements. Air quality data are visible in different ways depending on the pilot site and the collection method, as they can be visualized either as punctual measurements over a map (where each point corresponds to a specific sensor) or as an interpolated homogeneous map that estimates the pollution values in each spot of the city. Data can also be navigated in time, the date of the last available measurement varies in the different pilot sites, but in general data are gathered in a quasi-real-time way, thanks especially to the sensors acquired during the project.
- Census and demographic: basic demographic data such as population in the different neighborhoods, age distribution, ethnicity distribution, gender etc. are available for most of the pilot sites. This information is crucial to create important epidemiological statistics, as all the studied diseases have shown to have a

different prevalence and severity according to the demographic factors of the target population.

- Socioeconomic: Among the external factors that combined create the human exposome, socioeconomic factors play an important role, as they can largely influence quality of life and human behavior. For this reason, the PULSE WebGIS features several layers showing socioeconomic data such as poverty rate, crime rate, recycling, average income etc.
- Prevalence and incidence of asthma, type 2 diabetes, cardiovascular diseases and other correlated conditions are shown in different layers that can be navigated in space and time, in order to give a quick snapshot of the situation in the pilot sites concerning the diffusion of these pathologies, enlightening the most hit areas.
- Hospitalizations: besides prevalence and incidence, hospitalization rates for the different conditions are shown as well where available. Hospitalizations are another important public health indicator, as they reflect the tendency of the diseases to become severe, a phenomenon that is not necessarily strictly correlated to their prevalence.
- Various health indicators: for the pilot sites where the data is available, other generic health indicators are shown. This include specific risk factors, evaluation of the general health status of the population, wellbeing indicators, life habits. Examples of this kind of measures that can be found in the WebGIS are obesity rates, smoking rates, results of questionnaires about the perceived health status, happiness or depression indicators, physical activity etc.
- Environmental and urban factors: air pollution is not the only environmental factor that can have a noticeable effect on health and wellbeing. Other important factors that were considered in the creation of the WebGIS are noise, presence of green areas (parks, gardens etc.), percentage of land used for commercial or industrial purposes and traffic.
- Satellite images and maps: other environmental variables that are known to have an effect on health are monitored and shown in the WebGIS. In particular,

climatic conditions that affect air quality are shown mainly through two types of maps coming from satellite images: LST (Land Surface Temperature) images, that monitor heat waves and urban heat islands [47], and NDVI (Normalized Difference Vegetation Index) index images [48], that study the green coverage of the cities and the status of the vegetation, that correlates with a lot of factors (pollution, heat mitigation, moral status etc.).

Thanks to all the data represented, the PULSE WebGIS allows to have a comprehensive idea of the public health situation in the pilot sites, both through visual inspection and also allowing to add to the data a spatial description useful to analyze patterns and make predictions (as it will be shown in the next chapters). The ability to change the transparency of the layers and overcome them allows to visualize more indicators together and find possible visual correlations (e.g. overlapping air pollution maps and asthma hospitalizations maps it could be noticed that the areas where pollution is higher are the same where asthma tends to be more violent, or if this relation is not observed it could mean that there are other factors to consider). To facilitate data visualization and exploration even further, the PULSE WebGIS features two innovative systems:

- Side by side visualization of maps, that allows to easily compare more phenomenon, especially used in combination with the layer overlapping feature already described.
- A timeline that allows to visualize the data coming from a past period of time, and, where possible, to choose also the temporal resolution (daily, weekly, monthly, yearly data).

Figures 3.4 and 3.5 show two captures of the PULSE WebGIS: in the first one an image of New York where pollution and health data are combined is shown, with the side by side visualization functionality active; the second one shows a screenshot of Paris where air quality data from fixed monitoring stations are enlightened and the temporal bar is visible.

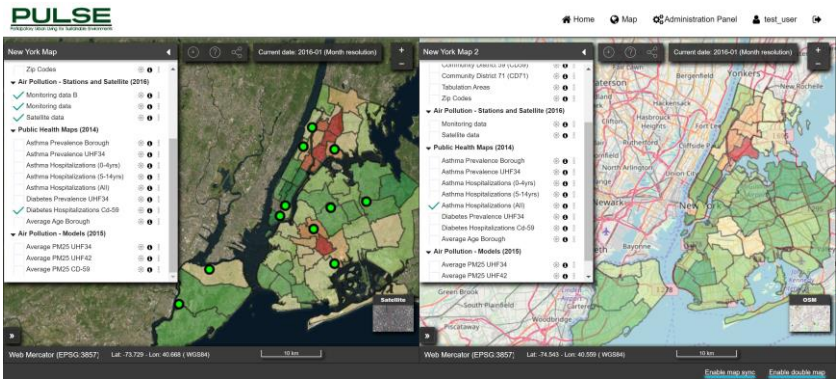


Figure 3.4: a WebGIS capture showing the side-by-side visualization feature on the city of New York. On the left, two layers are combined, one showing diabetes hospitalizations and one showing air quality measurements from the sensors (point data). On the right, a layer showing asthma hospitalizations is active.

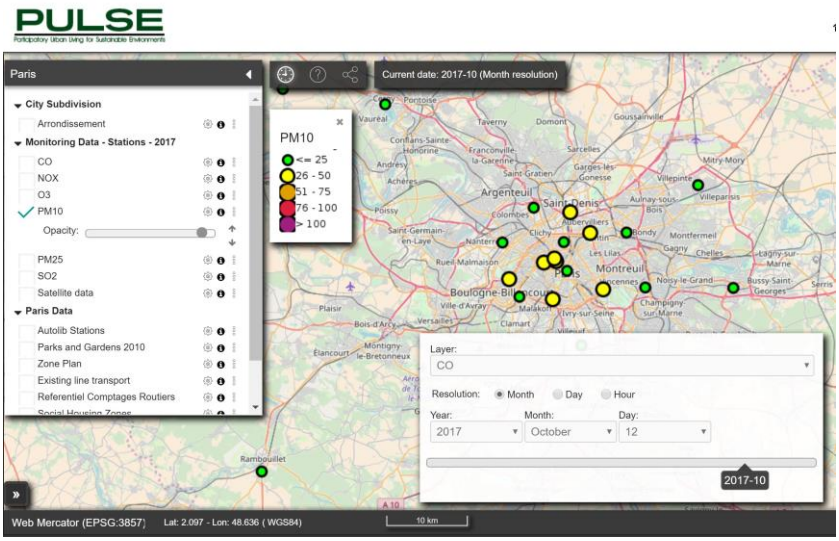


Figure 3.5: a WebGIS capture showing data from Paris. The active layer shows PM10 data taken from the punctual monitoring stations. The temporal bar that allows to navigate data in time is visible on the bottom right.

Some important applications of the WebGIS, in particular how it has been used for several parallel studies in the course of the project, are explained in chapters 4-7. It should be mentioned that the WebGIS is publicly accessible both by users and public health operators. All the information is aggregated and anonymized, sensible information that break the privacy rules of all the participating countries is not visible to anybody.

3.2.3. Back-end Infrastructure

PULSE has a complex architecture that includes a high variety of systems and tools that collect, analyze and visualize large quantities of heterogeneous data. One of the main components of this architecture is represented by the so-called *Back-end systems*, that represent the main engine of the project, as they store all the data collected from external and internal sources while they also allow for the integration and communication of all the interfaces (App, dashboards, FitBits, sensors, WebGIS). The Back-end systems are large and complex, a detailed description will not be given inside this dissertation as it is not fundamental to understand the aims and the results of the work presented, but the main systems can be categorized as follows:

- PULSE central database: the central database (PULSE DB) can be intuitively considered the main core of the PULSE system. Being PULSE a Big Data project, the necessary storage space for the database is very large. The PULSE DB is a relational database that contains all sorts of data, organized into several groups and categories in order to facilitate the data exchanges necessary for the functioning of all the systems. PULSE DB stores all the users data, where users are anonymized and identified only through an access code, and all the personal data are stored in other tables where they cannot be related to the account of origin. In particular, these data come from answers to questionnaires, FitBit integration, risk calculation and other App features. The PULSE DB stores also public health data, air pollution data collected by specific air

quality sensors and geographical information for the WebGIS.

- Health and Wellbeing models: a fundamental role in the back-end services and in the project in general is played by the risk models, that implement Big Data analytics and algorithms to compute risk factors of asthma, type 2 diabetes and cardiovascular diseases taking as input the data coming from the users (mainly from the questionnaires). All the users' App data are collected and stored in a dedicated place in the PULSE DB, and then used as input for these models, which are used and approved by the scientific community. The results of the models are stored as well and sent to the users via a set of dedicated Java services.
- Interface with FitBits and external sensors: another main functionality of these systems is the ability to interface a large variety of external sources of data, such as FitBits and air quality sensors of diverse kind. Thanks to an intricated combination of Java services and programs, different systems are integrated inside PULSE, and the data coming from them are usable for all the PULSE interfaces. This process is quite delicate, since there is a high level of heterogeneity in the hardware and software that has been integrated in the project, for example different kinds of air quality sensors have been used in the different pilot sites.
- Integration of external and internal tools: integration is at the basis of the PULSE concept, as the collected data and the components of the system are highly heterogeneous. The back-end systems contain important tools, Java methods and services to allow the integration of data and data sources. This includes both internal integrations, i.e. the data coming from different sources (Sensors, FitBits, App etc.) are all processed in order to be in the same format inside the database, and integration of external system, as some tools (e.g. personal exposure calculator and simulation tools) are developed in external platforms and the data are not retained inside the PULSE DB, but the back-end infrastructure provides all the necessary

functionalities that allow these systems and the central PULSE engine to communicate.

The back-end infrastructure has been mainly developed by the Serbian company BELIT, where the main servers are. The storage space of the PULSE DB is managed by the French company named Teralab, associated to the French partner of the project (IMT).

3.2.4. The PHO Dashboard

The PHO (Public Health Observatory) dashboard is another fundamental system of the PULSE architecture. Together with the WebGIS, it could be considered the most innovative and representative part, as it contains a lot of interactive features and it represents the connection between the citizens' wellbeing and the public health authorities' role which the idea of PULSE is partially based on. Even though it is usually described with the singular term, this system is actually made up several different dashboards, that together create a combination of tools, visualization portals, decision making aids and interactive programs that allow the users (represented by the public health authorities in this case) to assess the public health situation in the cities of interest and organize proper intervention strategies if and where needed. This set of tools is named *Public Health Observatory (PHO)*, as it can be considered as an observatory that allows to analyze the health and wellbeing situation in the city.

The PHO dashboard contains a lot of different features and integrates several external systems, some of which are created specifically for a subset of test sites, depending on the data available in each city. The main functions can be summed up as follows:

- Visualization: each pilot site has a dedicated dashboard where the user can visualize graphs, indicators, trends, plots and figures informing about several public health indicators related to the city itself. The main functionalities include visualization of aggregated health data concerning the treated diseases, visualization of wellbeing information, visualization of maps containing various phenomena and visualization

of air quality data and maps. No sensitive data revealing personal users' information are shown. With these tools, it is possible to have a quick idea of the general public health status of the city, concerning health, wellbeing and air quality in particular. It should be mentioned that, thanks to the spatial granularity of the data shown on maps, a neighborhood-level preview of the public health landscape is given to the public health operators.

- Analysis: The data that can be found inside the dashboard does not correspond uniquely to the data collected in the city. Besides knowing where the prevalence of asthma is higher or what is the air pollution status in the city, a public health operator can be interested in more complicated measures, for instance the probability of a certain phenomenon happening in the future or the percentage of population that can be potentially exposed to a specific risk factor. Some parallel studies have been performed during the curse of the project, most of which had the aim of studying correlations and interactions among the different data in order to obtain algorithms, predictions and procedures that could unveil interesting patterns or show measures that are not visible inspecting the raw data alone. Some of these studies are presented more in detail in the following chapters of this thesis, and they all represent an active way of exploiting heterogeneous urban data to predict public health outcomes in a heterogenous context.
- Simulation: public health issues in the urban environment can be addressed only through action, both from the citizens themselves and from the public health policy makers. This means that change is possible only through a dedicated urban planning strategy and active interventions. This process can be slow and expensive, therefore all the possible outcomes should be evaluated before undertaking it. To help with this issue, PULSE does not offer only visualization tools and results of analyses, it also applies all the knowledge gathered in the city to the creation of interactive tools that ease the intervention planning

process even further. These tools, named *Simulation Tools*, provide the policy makers with a set of models of the city with modifiable parameters that allow to simulate hypothetical scenarios and explore trends and possible outcomes of interventions. Detailed examples of these tools are provided in the course of this dissertation.

In conclusion, the PHO dashboard is an ensemble of tools and instruments that represents the connection between the citizens and the policy makers; thanks to the features of this system, urban public health is more understandable and decision making can be easier.

3.3. External Resources

In line with the definition of human exposome, PULSE contains an enormous quantity of data characterized by high variety. Although most of them are generated internally thanks to the App and the air quality sensors, many external sources of information have been involved in the creation of the system, especially to collect data regarding statistics about the population and socioeconomic landscapes. Furthermore, the involvement of external resources has encouraged the creation of parallel studies and projects, that addressed some specific public health issues in several cities following the trail created by PULSE.

This section reports a list of external resources used during the project, concerning both sources of data and parallel studies, with particular reference to a study performed in the city of Pavia, Italy, that is related to other topics that are presented in the next chapters of this dissertation.

3.3.1. Data Portals and Collaborations

There are basically two different kinds of external data that have been used in the project: data that can be found in the web and data gathered through specific collaborations. More in detail, the data on the web are usually categorized depending on whether they can

be freely accessed by anybody or they need to be purchased or obtained with a permission. Free web data are commonly called Open Data, and they are the simplest to obtain. The main advantage of open data stands in the possibility to collect them quickly without any bureaucratic process and without costs, but they are also often linked to some disadvantages concerning accuracy and granularity [49]. In particular, the absence of costs is often inevitably related to a limited use of resources to perform data cleaning and preprocessing, and this leads to a higher probability of having mistakes or outliers inside the dataset. Furthermore, open data is usually collected over the territory with a low spatial and temporal resolution, as data are usually available only for large districts and they are averaged over long periods of time (months or years), this is due both to the frequent lack of an appropriate amount of data to create statistically significant measures with high spatiotemporal granularity and to the fact that open data cannot contain anything that could potentially reveal sensitive information about someone in the city or violate privacy rules. Another frequent issue is the temporal lag that characterizes the data that can be found online with respect to the collection period, often due to long processing times. Despite all these problems, open data are becoming much more diffuse than the past in almost every city, as their importance to encourage research programs that can aid the public health panorama is recognized by the scientific community and the public administrations.

In PULSE, several sources of open data have been used both for the creation of maps and the database and for analysis and parallel studies. Some sources that are worth mentioning for the purposes of this thesis are:

- NYC Open Data Portal: this data portal [50] can be considered one of the largest open data repositories concerning only one city that is freely available online. It was created as the result of a cooperation between the city council, specifically the Mayor's Office of Data Analytics (MODA), and the Department of Information Technology and Telecommunication (DoITT). The idea behind the project is that open data available to everybody can help research projects and also inform the community and public or private agencies in order for the whole city to benefit from the

knowledge contained in it. This portal contains data of diverse kinds, spacing from health records (prevalences, trends, hospitalizations etc.) to environmental phenomena (pollution, green spaces, fauna etc.), passing from sociodemographic data such as education levels, ethnicity of the population, wellbeing indicators and much more.

- NYC Community Health Survey (CHS): the CHS is a telephone survey conducted annually by the Bureau of Epidemiological Services in the city of New York [51]. The survey is conducted on a random sample of about 10,000 people aged more than 18 that live across the city, and the questions have the aim to create a picture of the status of several health indicators in the city, concerning chronic diseases, personal behaviors and neighborhood status.
- NYC Data2Go Portal: this portal [52] integrates data coming from several different sources showing various socioeconomic and demographic variables, besides general urban statistics (e.g. safety, political tendencies, food habits etc.). The data are visible on a dashboard and can be downloaded without restrictions.
- New York State Department of Environmental Conservation: this public entity has a dedicated website where data concerning air quality measurements in all the official monitoring stations of the State of New York are available for download. The data is updated constantly, and quasi-real-time measurements are usually available, also with a high temporal resolution.
- 500 Cities Project: the data coming from this project has not been integrated in the PULSE database directly, but it has been used for a parallel study performed in the context of the project. The 500 Cities project [53] is a research program created by the CDC (Center of Disease Control) to collect a high quantity of data concerning health and life habits in the 500 most important cities of the USA. These data are collected with a high spatial subdivision. The dataset is better described in chapter 6.
- ARPA Lombardia air pollution and weather data: ARPA Lombardia [54] is a local agency that works for

the environmental protection of the region of Lombardy, located in northern Italy. This agency owns several air quality and meteorological monitors spread throughout the region, whose data can be easily accessed for research purposes through dedicated online portals.

Many other open data repositories have been used, all containing either pollution records or demographic/socioeconomic data, but they are not mentioned as they were not used in the work described in this dissertation.

Since the knowledge stored in open data can be limited for the aforementioned reasons, some data used in PULSE have been gathered thanks to the cooperation with local authorities and institutions. The main example is given by hospitalization data: hospital data are often sensible information and the open data portals tend not to show records with a high level of spatial and temporal detail. In PULSE, hospitalization records have been gathered and used in several occasions through private agreements, for example hospitalization records in New York City were obtained through a collaboration between the University of Pavia, the New York Academy of Medicine and the New York University that allowed to use the data stored in the SPARCS (Statewise Planning and Research Cooperative System) [55] dataset, under a contract that stated precise utilization rules. These data have been uniquely used for research and have not been shared with external parties or shown in the WebGIS or dashboards.

Other data have been collected through sources that are not available online, for example demographic and hospitalization data in Pavia, Italy, that has been gathered through a cooperation with the city municipality and the local health agency (ATS).

3.3.2. Low-cost Sensors Networks

As previously mentioned, in order to increase the granularity of the air pollution data gathered in the urban areas, several sensors have been acquired during the project. In detail, the PULSE architecture uses two different models of low-cost sensors: DunavNet [56] and Purple Air [57].

DunavNet is a Serbian company, founded in 2006, that proposes IoT solutions and architectures to create interconnected systems in different environments: farming, transportation, food industry, elderly care and environmental monitoring. One of the products they commercialize is the Ekonet sensor, a low-cost air quality monitoring station that measures all the main pollutants (PMs, NOx, SO2, CO, CO2) and weather variables such as temperature and humidity. These sensors are connected to a cloud system where the data are stored and can be easily visualized and analyzed. These sensors are relatively light and space-saving, therefore they can also be carried around. During the pilot phase of PULSE, some of these devices have been acquired, tested and integrated in the system in several cities, where they helped with the collection of spatially granular air quality data.

Another sensor model that was used in the project is the Purple Air PA-II sensor. This is a low-cost air monitoring device created by the American Purple Air company, it measures only PM1, PM2.5 and PM10 as pollutants plus temperature and humidity, but it has the advantage to be particularly small and easy to use, as it can be installed even by private citizens on balconies or external walls. The quality of the measurement is high if compared to other devices in the same price range. The university of Pavia has been the main utilizer of these sensors, as it deployed a high number of them over the city territory, thanks to the cooperation of the City Council and of private citizens. In detail, as of August 2020, 45 sensors have been deployed over a 70 km^2 territory, adding new monitoring stations to the two official ones already functioning and managed by the environmental protection agency (ARPA), that measure also the other main pollutants beside particulate matter (Figure 3.6). Pavia has developed an independent system related to this sensor network, as the raw data, although they are shared with the PULSE systems, are stored in the UNIPV servers where they can be accessed through SQL queries and Java methods. The latest measurements can be visualized also directly on the Purple Air website, through identification in a dedicated portal.

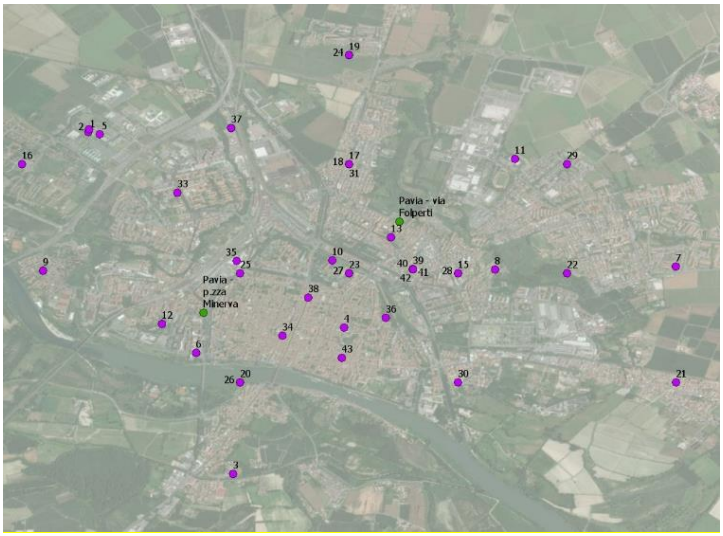


Figure 3.6: map of the air quality sensors currently functioning in Pavia. The PurpleAir sensors acquired specifically for the project are colored in purple, whereas the ARPA monitoring stations are highlighted in green.

Being the Purple Air PA-II a low-cost sensor (the price range is $200-$250 per unit), their reliability is expected to be lower than the official monitoring station used by the public agencies. In order to compensate, UNIPV has calibrated these sensors through calculation of the deviation of their measurements from the ones recorded by the ARPA sensors. From the practical point of view, a Purple Air sensor is periodically positioned close to an ARPA one and the measurements coming from both are compared for a certain period of time, and the Purple Air ones are adjusted with a linear regression performed with the parameters found by this comparison. This operation is better described in section 6.4. In general, the Purple Air sensors showed to have a high positive correlation with the ARPA ones (range 0.75-0.99), despite a small offset, as shown in Figure 3.7.

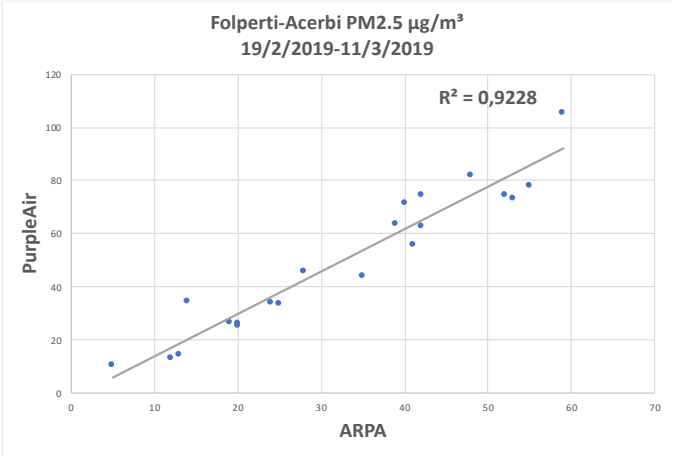


Figure 3.7: comparison of the daily averages measured by one PurpleAir sensor and the closest ARPA sensor. Correcting for a small offset, the correlation score was very high (0.9228).

The sensor network in Pavia has been used for different purposes:

- Establishment of a link with a dedicated App that can be downloaded and used by the citizens to be constantly informed about air quality in the city. The App is described in the next section.
- Correlation studies between air pollution and health outcomes, that have been performed thanks to the cooperation with the local public health agency (ATS), that has provided UNIPV with anonymized hospitalization data of the main hospitals in the province.
- Correlation studies between air pollution and climatic factors, especially wind speed and temperature. The calculation of these correlations has been useful in particular to study the effect of the 2020 lockdown related to the Covid-19 pandemic on air pollution, as reported in section 6.4.
- Creation of interpolated maps that allow to estimate the values of air pollution in every spot of the city. These measures have been used especially for the creation of

the personal exposure calculator, described in section 7.3.

The idea followed in PULSE, and confirmed in several studies, is that although low-cost sensors are generally less reliable than the expensive official ones, the possibility to buy and deploy them in large quantities has the advantage to allow to create better-defined air quality maps that consider also local factors. The possible lack of accuracy can be compensated using low-cost sensors in combination with the official monitoring stations.

3.3.3. The PULSE@PV App

One of the most important external resource used in PULSE is an App that constitutes the system created by the sensors network deployed in Pavia. This App works in parallel with PULSE, as it is not directly integrated in the main architecture, but it creates an extension of the benefits of the project for the citizens of Pavia, thank to special funds that have been granted to UNIPV by the local municipality. This App, named PULSE@PV works in a simple and intuitive way, and it has the aim to inform the citizens about air quality in order for them to be able to plan outdoors activities accordingly and be always aware of the risks. The main functionalities are:

- Visualization of the real-time PM10 value measured by the nearest sensor. This feature uses the positioning functionality of the user's smartphone, the value can be seen directly in the App's home page.
- Visualization of a list of sensors that were active during the last 30 minutes. Each sensor is marked with a feedback icon that indicates whether the measure corresponds to a safe air quality condition or it could lead to health risks.
- Visualization of the latest measures of PM10, PM2.5, temperature and humidity detected by all the sensors.
- Visualization of a map of all the sensors deployed throughout the city, color-coded according to the possible health hazard correspondent to the PM values detected.

Short sentences are provided together with color-coding to explain to the users the risks involved in the exposure to the air pollution levels measured. An informative page that explains the project, the aims and the data use can be accessed by each user. This App is more effective if used in combination with the Pulsair App, that also gives health-specific feedbacks. As of August 2020, PULSE@PV has been downloaded by approximately 200 users, using both Android and IOS operative system. Figure 3.8 shows some screenshots of the App.

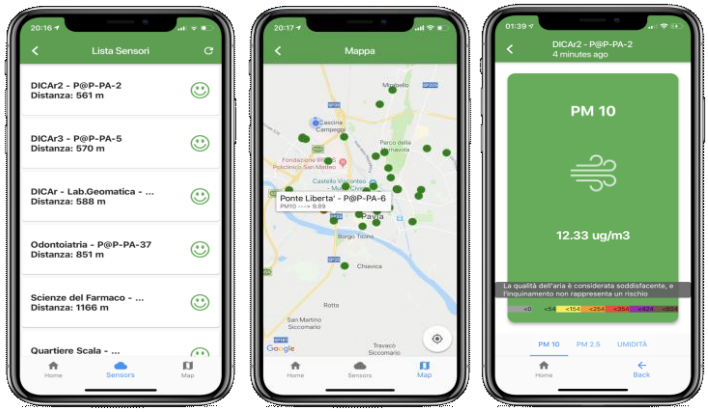


Figure 3.8: Screenshots of the PULSE@PV App. From the left: the list of sensors ordered by distance with a feedback icon, a map of the sensors, the most recent measurement of the nearest sensor with an accompanying sentence that provides the danger level associated to the pollution status.

Chapter 4

Spatial Analytics

As it has been written several times in the first chapters of this dissertation, the importance of the spatial dimension has been one of the pivotal ideas of PULSE, its development and its applications. When it comes to studying public health, geography plays a central role in all the phases of data management, from collection to visualization and analysis.

In this chapter, a more detailed description of how space has been added to several features of the system is given. In particular, after an introduction on the main concepts, the description of some of the main spatial analytics is reported.

4.1. Spatial Enablement

The concept that stands at the basis of the importance of the spatial dimension in PULSE is named *Spatial Enablement*. This term indicates the ability to add geographic information to the data collected in a certain timeframe and place to improve analytical and visualization capabilities for data analysis [58]. The idea of including spatial information on the data is not new, as most of census, socioeconomic and environmental data have always been collected with some kind of spatial reference. For example, census data refers to neighborhoods, and hospitalization data refers to hospitals that are located by specific geographic coordinates.

Knowing where a certain phenomenon happened, besides the magnitude of it, allows to characterize certain public health patterns with more precision, and in some cases location can be

considered crucial, for example in epidemiological studies of viral diseases, where knowing where the virus is moving is fundamental to decide how to intervene on it.

Although this concept is not new, the past research shows that the spatial dimension is still not always exploited in a lot of urban public health studies, in particular, during the activities of creation and development of PULSE, the consortium found mainly two issues:

- A lot of importance to space is given when it comes to visualizing the data, but not in their analysis: several geographic machine learning-based tools have been developed in the last decades, but their use in urban public health is limited.
- When the spatial dimension is considered in urban public health, the spatial granularity of the studies is often insufficient to create significant statistics that assess all the issues of a city. Neighborhoods are usually grouped together hiding important local data that could be informative [59].

PULSE tried to give a possible solution to these problems, creating a system that increases the spatial granularity of the collected data and exploring new ways to apply spatial enablement in urban public health. A quick explanation on how this is done is given in the next sections.

4.1.1. Visualization

The most obvious application of the concept of spatial enablement is the ability to visualize geographical patterns and spatial details of various phenomena. The PULSE WebGIS is a clear example of this, as it allows to visualize maps showing how several public health variables are distributed throughout the urban environments analyzed in the project.

Proper visualization of public health phenomena is the first step to understanding important elements of the health panorama of a certain location, that leads to taking informed action to control the spread of dangerous phenomena. The most typical example of this is the observation of the propagation of contagious diseases, but

this is true also for all the other kinds of public health data. In urban health, for instance, a different distribution of incidence of a certain disease or an unequal distribution of socioeconomic indicators could highlight that some neighborhoods face unhealthy conditions more than others, and the comparison of data through time could explain whether these conditions tend to be spreading or contained in the city. Moreover, visualization of multiple public health phenomena can help finding correlations and links, for example air pollution and prevalence of respiratory diseases or calls to the emergency services.

To this end, all the innovative features of the PULSE WebGIS introduced in chapter 3 contribute to the creation of a proper visualization tool that allows to aggregate data and maps to have a comprehensive vision of the public health panorama of the city. Besides open data and satellite images, also the data gathered inside the project itself are spatially enabled, as users agree to share their zip code of residence or to be tracked by the GPS functionality in their phones, thus geography is always added to the collected data.

4.1.2. Analysis

Visualization is not the only way in which spatial enablement can become useful to understand public health phenomena. The rise of awareness on the importance of geographic information in the detection and prevention of dangerous phenomena (not limited to public health) has brought also to the development of new analysis techniques based on spatial integration. All the modern software used for GIS creation and treatment nowadays contain a large set of data analysis tools that are able to perform operations such as clustering, similarity research, mapping, classification, data fusion etc. including the spatial dimension as element in the equations that define the methods.

One clear example of this is a class of methods named *Spatial Clustering* [60], that are designed to find groups and patterns of similarity in a dataset similarly to the traditional clustering methods, with the addition of weights or equations that tend to cluster together data that are collected in areas geographically close to each other. The same idea can be applied also to supervised learning, generating the so-called *Spatially Weighted Classification*

[61], i.e. a set of classification techniques where spatial weights are introduced in order to modify the distance metrics used taking into account also the geographic distance of the observations. This is particularly useful to classify sensors data or to analyze data that are significantly dependent on space.

Spatially enabled analysis tools are used also for observatory analyses, with modified statistical methods that perform distribution analysis and statistical tests introducing parameters that give information about the geographical elements of the features, such as geographical mean, centroid and orientation. The creation of these methods has led to the raise of a new field of research named *geostatistics* [62], i.e. a set of statistical methods used to predict or analyze values associated with spatial or spatiotemporal coordinates. These coordinates are introduced in the methods themselves. Geostatistics has a very wide range of applications, in different fields such as the mining industry (to identify the distribution of minerals and materials in order to optimize the mining process), environmental sciences (to quantify pollutants and predict their distribution), meteorology etc.. Applications in public health are relatively new. The main rationale behind the idea of geostatistics is that geographical phenomena are usually mapped taking samples in distant areas, creating the need of data integration and generation of predictions, with the related uncertainty measures.

4.1.3. Georeferencing

One of the most important actions in the analysis of geographical data is georeferencing. This word is referred to the situation in which a set of coordinates of a map or an image is somehow remapped to a geographic set of coordinates, with an identifiable correspondence between the two sets. In other words, thanks to georeferencing, the information contained on a map or digital image can be associated to its real location in the world.

This process is pivotal in spatially enabled storage of information and analysis, as it allows to make aerial and satellite images useful for mapping real phenomena and it explains how data relate to the imagery.

Besides this, georeferencing allows to perform important operations to overlap, analyze and integrate information. For

example, different images could contain different useful information coded in different ways, georeferencing allows to unify this information in order to perform analysis or visualization of phenomena of the integrated data. Also, information coming from punctual locations, such as sensors located in space, or coded in different ways as coming from different sources can be easily integrated and related to the real world. Also temporal analysis can be fostered by georeferencing, as sometimes different images or maps could contain data referred to different timeframes, so their combination can be revealing of trends and patterns of the studied phenomena depending on the temporal dimension.

There are several ways of georeferencing an image or a set of data, most of which nowadays are already integrated in the modern GIS tools, through which georeferencing is possible for a set of points, lines, polygons, images or even 3D structures. The only necessary requirement is that the georeferencing method uses a unique identifier that express a correspondence with one location. Some data are already georeferenced depending on the way they are collected, for example GPS tracking devices record latitude and longitude, actually georeferencing the collected points.

Georeferencing of images can be a more complicated process, as some control points with known geographic coordinates have to be set, with a pre-set coordinates system and established projection parameters [63]. Sometimes images are encoded using specific GIS file formats or are accompanied by a world file, i.e. a conventional text file format with 6 lines containing information on the transformation to be performed to associate a geographical location to each pixel of the image.

4.2. Spatially Enabled Methods

As explained in section 4.1.2, spatial enablement is frequently used to perform data analysis for data with an important geographic description. A lot of Machine Learning and statistical techniques have been adapted to include the spatial dimension in the analysis, usually with the addition of weights and parameters that give a crucial importance to the location where the data have been collected. In the next two subsections, two spatially enabled methods that have been used in the work reported in this thesis are explained.

4.2.1. Spatial Clustering

The aim of all clustering techniques is to group together the objects of a dataset into a series of subgroups based on their similarities [64]. Spatial Clustering algorithms follow the exact same principle, grouping together objects in subclasses such that similarities between objects in the same group are maximized and those with objects in the other groups are minimized, but with datasets that have a clear spatial reference that is taken into account in the similarity search process. Currently, spatial clustering is widely used in the field of spatial data analysis, such as land use studies [65], earthquake analysis [66], geographic customer segmentation [67] and public health [68].

Spatial Clustering algorithms can be roughly divided into seven groups: partitioning, hierarchical, graph-based, grid-based, model based and combinational algorithms [60]. The choice of using one type of algorithm over the other depends strictly on the problem to solve, since some provide more precise information on the position and shape of the clusters and the geometrical properties of the attributes than others.

The most common spatial clustering algorithms are usually available as tools on the main GIS softwares, such as ArcGIS [69]. ArcGIS, developed by ESRI [70], is one of the most famous GIS analysis and elaboration programs, as it contains several tools for the upload and the analysis of GIS layers, orthophotos, raster images, etc.

This software features a large spatial statistics toolbox that includes several spatial data analysis algorithms. One of them is the *Grouping Analysis* tool, that performs spatial clustering with algorithms and parameters that depend on the settings gave as input by the users [71].

In particular, when options on the spatial constraints are specified, the algorithm employs a connectivity graph to find natural groups, otherwise the tool uses a classical K-means algorithm.

This tool takes point, polyline, or polygon Input Features, a unique ID field, a path for the Output Feature Class, one or more Analysis Fields, an integer value representing the Number of Groups to create (if known), and the type of Spatial Constraint -if any- that should be applied within the grouping algorithm. The output is a new Output Feature Class that contains the fields used I

the analysis and a new field indicating the group each feature belongs to.

When the number of groups is not known, the algorithm tries to find the optimal number of groups using the Calinski-Harabasz pseudo F-statistic, i.e. a ration reflecting the within-group similarity and the between-group difference:

$$\frac{\left(\frac{R^2}{n_c - 1}\right)}{\left(\frac{1 - R^2}{n - n_c}\right)}$$

Where n is the number of features, and n_c the number of clusters. The parameter R^2 is a score that reflects how much of the variation of the original data was retained after the clustering process, and is calculated as:

$$R^2 = \frac{SST - SSE}{SST}$$

Where

$$SST = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} (V_{ij}^k - \overline{V^k})^2$$

$$SSE = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} (V_{ij}^k - \overline{V_i^k})^2$$

Where n_i is the number of the ith cluster, $n_v$the number of variables used for the clustering process, $\overline{V^k}$ the mean value of the kth variable, $\overline{V_i^k}$ the mean value of the kth variable in the ith cluster and V_{ij}^k the value of the kth variable of the jth feature of the ith cluster. Summing up, SST is the total sum of squares of the differences between the mean value of each variable of each feature and the mean value of each variables, and therefore it represents the between-clusters differences, whereas SSE is computed in the same way but using the mean of each variable for each cluster, and therefore represents the within-cluster similarity. For this reason, the higher R^2, the

better the discrimination among features with the selected number of clusters.

If the spatial constraint option is activated, the Grouping Tool clusters together only polygons that share an edge or a vertex, in alternative the user can input a weights matrix that assigns an initial score to cluster together areas that are known to have similarities a priori.

The algorithm which the spatial clustering is based on is the *minimum spanning tree* [72], i.e. when the spatial constraint is activated, the algorithm designs a graph that maps the relationships among the features in the neighbor areas: each feature becomes a node in the tree that is connected to other nodes by weighted edges, where weights are proportional to the similarity between the objects. The tree is then pruned in a way that minimizes the dissimilarity in the resulting groups, obtaining a final minimum spanning tree.

4.2.2. Geographically Weighted Regression

Spatially enabled data analysis methods, besides clustering, also include a set of supervised learning methods and other statistical analysis tools, such as geographically weighted classification [73] and geographically weighted regression (GWR) [74]. This last method has been used in a study reported in this thesis in chapter 6, therefore it is presented in this section.

The concept that stands at the basis of GWR is the same of regular regression. Linear regression is usually represented in the form:

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$$

Where **Y** represents the vector with the measurements of the dependent variable, **X** the vector with those of the independent variables, **β** the coefficients and **ε** the error terms. The aim of the regression is to find a relationship considering each couple of points (x_i,y_i) such that the variable **Y** can be predicted knowing the values of **X**. In mathematical terms, this is done by estimating the coefficients in a way that minimizes the error between the observed values of Y and those obtained with the model $\mathbf{Y}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$. This could be written with the formula:

$$\widehat{\boldsymbol{\beta}} \coloneqq min \sum (y_i - (\boldsymbol{X\beta})_i)^2$$

That indicates how the coefficients are estimated minimizing the squared norm of the difference between the observed values and the ones calculated by the model. A weight could be introduced in this formula to give a different relevance to each observation, modifying the equation as follows:

$$\widehat{\boldsymbol{\beta}} \coloneqq min \sum w_i (y_i - (\boldsymbol{X\beta})_i)^2$$

GWR works in this way, calculating a weighted linear regression with weights that give more relevance to observations that are geographically close to each other. In the case of the GWR algorithm we used, we created a grid of regular spaced dots that overlapped the GIS layers we had (see chapter 6 for details), and we computed a different linear regression for each dot. Our GIS layers were described in polygons, so we localized the values of each observations related to each polygon in its centroid. This allowed to compute a set of linear regressions, one for each dot, where the values of the observations related to the polygons where the same, but the weights changed. In our case, we created the weights as follows:

$$w_i = e^{-\frac{d_{ij}^2}{s^2}}$$

Where d_{ij} represents the distance between the ith dot and the jth centroid, and s is an arbitrary selected threshold. As reported in chapter 6, our threshold was set to 5 km and the dots were spaced all 1 km from each other.

4.3. Convolutional Neural Networks

4.3.1. The Main Idea

The word *Artificial Neural Networks* refers to a class of algorithms based on the connection of multiple nodes (artificial neurons), inspired to the anatomy of the brain, composed by several synapses connected together [75]. Each artificial neuron takes one or more inputs from other neurons or external sources and transmits an output, deriving from an elaboration of the inputs, to the other neurons it is connected to. The transmission is usually regulated by a non-linear function and by weights. Neurons are usually organized into layers. The layer with the neurons that receive external data is named *input layer*, and it is connected to the *output layer*, i.e. the layer producing the output of the analysis through a set of *hidden layers*. The neurons of each layer are connected exclusively to the ones of the preceding layer, from which they receive the inputs, and to the ones of the subsequent layer, to which they transmit their output. Figure 4.1 shows a representation of a typical architecture of an artificial neural network.

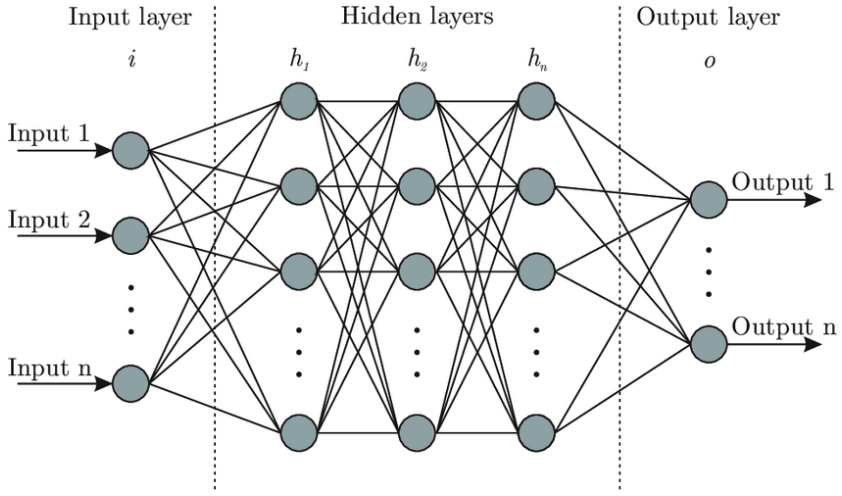


Figure 4.1: example of the architecture of an artificial neural network.

Among the main elements of a neural network, one of the most characteristic is the *propagation function*, that computes the input of a neuron as a weighted sum of the outputs of its predecessors. In this phase, *backpropagation* [76] (i.e. an algorithm used to adjust the weights of the network based on the error detected during the learning process) plays an important role in improving the accuracy of the learning results. Another element that could be present is represented by the *hyperparameters*, i.e. a set of fixed parameters that are set at the beginning of the learning process and remains constant during the process (e.g. number of hidden layers, learning rate etc.). Artificial neural networks have many applications, they are usually applied to machine learning problems both supervised and unsupervised, since they can learn information from examples thanks to the propagation of information within their neurons. Other applications include systems control, video gaming, pattern recognition, face identification, translation and many others.

An important class of neural networks, that is gaining more importance in the last years, is the so-called *Convolutional Neural Networks*. This is a class of neural networks belonging to the deep learning paradigm [77], as the networks are formed by a high number of layers. These layers are usually fully connected, i.e. each neuron in one layer is connected to all neurons of the following layers. The word "convolutional" derives from the convolution operation [78], that is performed in most layers and stands at the basis of the network, as it transforms an image into a feature map. Another important characteristic of most CNNs is the presence of *pooling layers*, that reduce the dimensions of the data combining clusters of neurons together.

4.3.2. Image Analysis

CNNs are very versatile algorithms that can be tuned for numerous different problems, as all neural networks, but they are typically used to perform the analysis of images and 2D objects, including videos. They are mostly applied in pattern recognition and image classification, leading to a large spectrum of possible applications, from medical analysis (e.g. support to the diagnosis through analysis of medical images) to forensics (e.g. face recognition) and automated driving (sign recognition).

Several applications of these algorithms have been reported also regarding satellite images analysis and geographic/geomatics applications. For instance, CNNs can be used to monitor land cover changes, useful in environmental planning or disaster mitigation [79], or even to identify possible environmental hazards such as oil spills [80]. Applications in the urban environment can be found as well, for example Cao et al. [81] proposed a pipeline based on CNNs to classify aerial and street view images of New York City for urban planning and management.

The main problem of CNNs is that their training usually requires vast datasets with numerous examples, that are not always easily available. One solution to this issue is to use pre-trained networks to create feature maps of images different from the kind of images used for the training. This operation is based on the concept named *transfer learning*, i.e. the attempt to solve a problem using a method originally created to solve a different, even if similar, problem. The study reported in section 6.3 for example shows the application of a pre-trained CNN used to classify artwork for the creation of feature maps of satellite images of urban areas.

Chapter 5

Interactive Simulation Tools

Improving public health in an urban environment can be a difficult challenge that involves public health authorities and city municipalities, besides citizens themselves. Spatial analytics such as the ones presented in chapter 4 can help analyzing the health problematics of the city, finding the neighborhoods that mostly need interventions and designing the best plans to improve the public health panorama of those neighborhoods, but interventions on factors such as urban structure and education of the citizens require time-consuming and expensive processes. For this reason, the use of other technological tools that allow to better predict the effects of urban planning strategies, in addition to the ones presented on chapter 4, can help making the intervention processes easier and safer.

To face this issue, the PULSE dashboard features a set of *simulation tools* built to respond to urban planning related questions and predict possible scenarios deriving from specific actions.

In this chapter, the methods used to create these tools are briefly described.

5.1. Agent-Based Models

PULSE's simulation tools are based mainly on the modeling paradigm known as *Agent-Based Modeling*. Agent-based Models (ABMs) are a class of very versatile tools that can be used to model a vast variety of phenomena, based on the simulation of the

interactions of several entities located in a shared environment. As in other scientific environments, the term *model* contained in the ABM definition refers to an abstraction or representation of a given reality [82]. A simulation model refers to the algorithms, mathematical expressions and equations that encapsulate the behavior and performance of a system in the real world scenarios, as defined by Abar et al. [83].

In this subchapter, after a more detailed presentation of agent-based paradigm, some scientific applications are reported, followed by their extension to public health and the subsequent integration into the PULSE system.

5.1.1. Simulation of agents' interactions

The key to the ABM simulations stands in the concept of *Agent*. An agent can be thought as something similar to the informatics concept of object, i.e. an element that incapsulates attributes, methods and operations of a software module, but it presents some crucial features that make it a different entity. Differently from an object, the agent is not expressed as a container of attributes and methods, as it contains a higher level of abstraction that allows it to be expressed in terms of its intended actions [83]. An agent is an autonomous and independent element, capable of performing actions by itself. In the definition of an ABM, the user sets a number of variables and mathematical relations that define the characteristics of a shared environment in which the agents operate and the interactions the agents can have among themselves and the surrounding environment. Agents are intuitive, they have the ability of perceiving all the changes in the surrounding domain and autonomously respond to them [84]. Summing up, it could be said that an agent is an independent entity that acts in relation to its own beliefs and behavioral characteristics, reacting to the surrounding world and to other agents.

The ensemble of a high number of agents in a shared environment that creates an abstraction of a real-world phenomenon, either observed or created, generates the Agent-based simulation paradigm. From a practical point of view, an ABM can be considered a computational model based on a set of action-reaction protocols and mathematical relations that create a dynamic environment in which agents operate, in order to observe deriving

insights on the emerging behaviors or possible changes in the environment. The most diffuse use of this paradigm is the simulation of realistic scenarios with a set of self-governing agents that represent living entities that act with behaviors and beliefs similar to those of the real-world entities they represent. Agents can represent all sorts of real-world entities, either animated or inanimate, such as people, animals, plants, objects, vehicles etc., as long as they possess intrinsic behavioral tendencies and are able to actively interact with the environment. In most ABMs, there are different categories of agents, each one with different characteristics and behaviors, that interact with each other. Sometimes agents within the same model represent the same real-world entities, but they are subdivided into subcategories, named *breeds*, according to some peculiar features (e.g. whether they are sick or healthy).

ABMs can be programmed using specific software, such as NetLogo [85], or using dedicated toolboxes of existing programming languages, for example the Python package MESA [86].

5.1.2. Applications

Considering that agents can model a large variety of entities, ABMs are extremely versatile and can be used in a large variety of science fields, such as epidemiology, social sciences, biology, economics, finance, meteorology etc.

The field of application depends solely on what the agents represent, for example in an epidemiological model agents can be people (possibly divided into sick/contagious and healthy) or viruses, whereas in a biological model they can be an abstraction of cells or microorganisms.

In general, ABMs can be used to study phenomena where there is an important cause-effect mechanism, such as in epidemiology [87]. In this field, ABMs are a useful tool to simulate the outbreak of a disease or the progression of a contagious pathology that can be passed from one agent to the others. Many ABMs that simulate the outbreak or the progression of influenza clusters have been created in order to find new perspective for prevention or intervention and to study the effect of vaccination campaigns [88]–[90]. The applications of these study vary from understanding the

burden of a certain disease in hospital systems to studying the social implications of the epidemics, or even understanding the most important vectors of disease transmission, as it was done by Cooley et al. [91], who studied the role of subway travel in the spread of an influenza epidemic in New York City and concluded that only about 4% of the contagions happened in the subway, thus concentrating the interventions on the transportation system could be a relatively ineffective strategy.

ABMs can be applied also in fields where people are not directly involved, such as biology, for example Gorochowski et al. [92] used ABMs to model bacterial populations, using agents as an abstraction of bacteria rather than people.

Agents can also be inanimate objects, as long as they possess autonomy and specific behaviors. For example, there are ABMs that study air traffic where agents are the abstraction of airplanes [93] or ABMs that study the chemistry-related self-organization properties of some substances where agents are molecules [94].

These examples are only a glimpse of the enormous range of phenomena and dynamic systems that can be studied using this technology.

5.1.3. Extension to Public Health

Being a multidisciplinary field, Public Health offers a large spectrum of possibilities of applications for ABM technology. Among other things, public health is also about interventions, problem solving and active actions, all elements that are at the basis of an ABM. ABMs in public health can be used to simulate health-related scenarios, answer to "what-if" questions and plan interventions to mitigate health risk or increase wellbeing in a certain environment. Although the application of ABMs to public health is relatively new (the first studies appeared within the last 10 years), one could say that all the ABMs developed to study epidemiology, economics, climate etc. are somehow related to public health, as solving problems in these fields contributes to the global health and wellbeing status of the population.

Historically, ABMs in public health have always been used to create epidemiological models of infectious diseases [95], as it seemed to be their most natural application. With time, they started to be used also to study health-related human behavior, for example

some models were developed to study the prevalence of smoking in the population, for example by investigating the influence of social factors on smoking [96] or the effect of the introduction of e-cigarettes on the smoking behavior [97]. Smoking is not the only behavior that has been simulated with public health ABMs, for example Yang et al. [98] investigated the population's tendency to walk in an urban environment related both to external factors (i.e. the status of the neighborhood) and to their personal characteristics, including ability to walk, experience, age etc. Other behavioral studies have been performed with ABMs, for example Auchincloss et al. [99] studied the impact of social segregation to different dietary habits that can lead to an increased risk of obesity or certain illnesses.

Fundamentally, ABMs in public health can be a powerful tool to explain or predict health outcomes, providing insights on the cause-effect mechanisms at the basis of the rise of health-related problematics and behaviors. Following the same idea, they can also be used to conduct virtual experiments of interventions and policies to reduce the burden of certain diseases, as pointed out by Tracy et al. [95].

Unfortunately, the use of ABMs in public health presents some limitations as well, that can be mostly identified in three points: finding the best trade-off between model simplicity and model realism, the possible effect of confounders in the definition of the relations between the modeled objects, and the lack of simulation of the steps of the interventions. In detail, these problematics can be explained as follows:

- When building a model, one of the main rules to follow is to try to keep it as simple as possible [100], but the simulation of a public health interventions scenario requires a certain level of detail in the representation of the real-world system, in order to generate meaningful results on the possible impact of the health intervention. The balance between these two elements is not always so easy to find, and trial-error strategies are often used to tune the model properly.
- Sometimes it could be difficult to model all the relations that are needed to create a reliable abstraction of the real-world system represented by the model. This difficulty is usually a direct consequence of the

lack of empiric data to model a certain phenomenon. When empiric data are actually available, sometimes they could be taken from observational studies performed on populations with different characteristics and casual patterns than the modeled one, thus introducing confounders in the model [101], [102]. Furthermore, validation can sometimes be impossible as it would need other data taken independently from the ones used for the creation of the model. This is a known problem in Public Health Agent-based Modeling, but on the other hand even an imprecise ABM can provide useful insights on the mechanisms at the basis of the development of a health situation.

- The simulation of interventions is usually performed under the assumption that specific actions could lead to a certain percentage of success, but the steps that would be required to obtain this success are rarely considered. For example, a classical public health ABM could simulate the effect of a 10% or 20% reduction in the obesity rate of a city on the prevalence of diabetes, but it's usually difficult to gather enough data or to build a model complex enough to simulate all the steps that would be really necessary to come to such reduction.

In spite of these limitations, if properly used by taking into account their potential together with their limitations, ABMs are powerful tools for public health, as they are useful to model a wide variety of interactions, behaviors, cause-effect paradigms and environmental and social phenomena.

Geography can represent an important addition to these kinds of models, ABMs are in fact usually created a simulated generic environment, as they focus on the agent's interactions rather that the real-world setting of the scene. In several public health scenarios however, the location of the studied phenomena can be of crucial importance, especially in urban planning. With time, several tools to integrate GIS data inside the Agent-based world have been developed and used [103], [104].

5.1.4. Integration in PULSE

The PULSE dashboard, presented in section 3.2.4, includes a collection of observational and interactive tools for the public health policy makers through which they can explore the city data, observe phenomena, study health related problems and find tools to ease the intervention planning process. Some of these tools are simulation tools developed using the Agent-based paradigm.

During the design of the project, ABMs were identified as a proper technology to develop innovative simulation tools to model public health-related scenarios using the large quantity of data gathered in the context of the project. In PULSE, ABMs are concepted as a tool mainly to simulate "what-if" scenarios, to study the impact of the possible changes in some variables of the public health panorama of the city and to generate insights on the possible effects of intervention strategies.

ABMs in PULSE are implemented creating state-of-the-art public health models, that integrate data observed in different studies and incapsulate GIS technology to make the simulation environment real. The models are developed using the software NetLogo, specifically designed to create ABMs, and are integrated inside the PULSE dashboard, using a Java-based embedding feature that NetLogo provides.

So far, two simulation models have been created to study two different phenomena, both related to PULSE and the data gathered inside it: the influence of air pollution and socioeconomic factors on the asthma hospitalizations rate in New York and the effect of traffic and wind speed on the air pollutants concentration in Pavia, Italy. These models are described in detail in chapter 7, subchapters 7.1.1 and 7.1.2. It should be mentioned that both these models integrate GIS technology in order to perform the simulations over a real-world environment made of streets, sidewalks, parks, buildings etc., and to give the observer a more realistic idea of what could be the effects of a variable change in the health panorama of the city.

5.2. Multi-layer urban traffic modeling

One of the most typical problematics of every urban environment is notoriously traffic. Elevated traffic levels can influence health and wellbeing in several ways [105], for example increasing pollution and noise and affecting also mental health of the population [106]. Traffic is a widely studied and modeled topic, as there are numerous studies that have traffic modeling as first aim, with the global idea of finding a way to reduce the probability of traffic jams and congestions that can be dangerous at several levels. Traffic modeling is of primary importance also for public health, especially in the urban environment, not only for its strict correlations with pollution- or mental health-related outcomes, but also because it is a phenomenon that can be observed and intervened on relatively easily, although traffic modeling is known to be difficult.

Among all the tools that have been used to model traffic, ABMs appear in several occasions [107], [108], as they have been identified as a proper instrument to simulate how traffic flow would change if some interventions were applied, for example closing roads, reducing the amount of cars, changing the traffic lights patterns etc.

In an urban public health project such as PULSE, traffic cannot be neglected among the public health variables that were chosen in the models developed. During the project, the technical team of the Laboratory of Biomedical Informatics M. Stefanelli of the University of Pavia started a collaboration with the Laboratory of Dynamic Systems of the same university, that includes a team specialized in traffic modeling. Although traffic has been widely studied in the world of simulation tools, as it has been the relation between pollution and health, comprehensive simulation models able to determine the health status of the population in dependence of a large variety of factors are generally rare even in the public health environment. For this reason, in PULSE we decided to unify different competences and create a large simulation model that incapsulate both the dynamic system-based traffic modeling and the spatially enabled health analytics, creating a large Agent-Based simulation tool where agents, representing people, are exposed to a complex combination of factors (environmental, socioeconomic and demographic) whose their health depend on. This model is

described in chapter 7, section 7.2, but the underlying traffic analytics are described in this paragraph.

Generally speaking, there are three kinds of approaches in urban traffic simulation modeling: the macroscopic approach, the microscopic approach and the mesoscopic approach [109]. In short, the term *macroscopic* refers to models that represent traffic flows in an aggregate way whereas the term *microscopic* refers to models where the interactions of individual vehicles are considered. *Mesoscopic* models are a sort of middle ground, as they share properties from both models [110].

Unfortunately, extensive traffic data could not be gathered during the PULSE project, as their availability as open data is extremely limited and they are often difficult to obtain even by purchasing them. For this reason, we opted for a macroscopic approach, modeling traffic in an aggregated way in the different neighborhoods of the city, and adding its effects to the health risk of the population computed with spatially enabled models such as the GWR. In particular, our reference model is the one adopted and presented by Menelaou et al. [111], that built a regional-level model that combines route guidance with demand management, based on the origin and the intended destination of the vehicle flows inside the network. Basically, the city is divided into regions, and the traffic flow of each region is described using the *Network Fundamental Diagram* (NFD), that describe in a relatively simple way the macroscopic relations between the three main mobility patterns, i.e. speed, flow and density [112]. This diagram is composed by two different regimes, i.e. the free-flow traffic regimes where traffic flows at maximum speed and the congested regime, that occurs when there is a congestion due to traffic reaching a certain density inside the considered region. Figure 5.1 reports an example of such diagram.

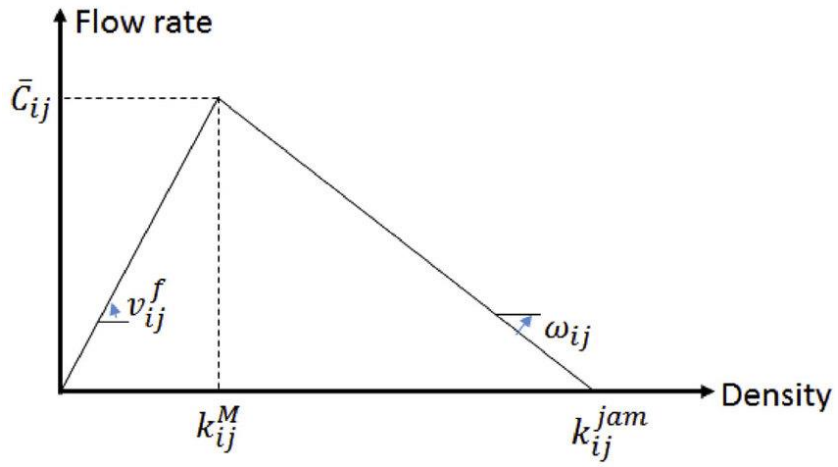


Figure 5.1: Example of a triangular Network Fundamental Diagram. Source: Ma et al., "Emission modeling and pricing on single-destination dynamic traffic networks".

The model implemented is defined by a set of different parameters. First of all, let the urban area be divided into R different regions denoted by $r \in R = \{1, \dots, R\}$. The traffic flow in each region can be described by a triangular NFD. Let also $O \subseteq R$ and $D \subseteq R$ be the regions considered as origins and destinations respectively, and $J_r^- \subseteq R$ the set of neighboring regions to $r \in R$. Similarly, $J_r^+ = J_r^- \cup \{r\}$is defined such that:

$$J_r = \begin{cases} J_r^+ \text{, if } r \in D \\ J_r^- \text{ otherwise} \end{cases}$$

Also, let $d_{od}(k)$ be the number of new vehicles that request to enter the region $o \in O$ to go to the region $d \in D$ at the time step k, i.e. the so-called external demand. The admitted external demand, i.e. the vehicles that actually enter inside the region is indicated with $\tilde{d}_{od}(k)$, and is determined by the physical limits of the regions, the maximum possible demand that can enter a region at a certain time step, denoted by D_{od}^{MAX} and demand management rules that allow only a portion of the vehicles that request to enter the region to actually access it. The remaining flow that request to enter a certain region at a certain time step can be defined as:

$$D_{od}(k+1) = D_{od}(k) - \tilde{d}_{od}(k) + d_{od}(k), \qquad k = 1,2 \ldots,$$

At this point, we can define the variable $\rho_r(k)$ as the density, expressed in vehicles/km, of a region $r \in R$ at the time step k, and $\rho_{rd}(k)$ the portion of this density that is destined to the destination $d \in D$ as follows:

$$\rho_r(k) = \sum_{d \in D} \rho_{rd}(k)$$

$$\rho_{rd}(k) = \sum_{j \in J_r} \rho_{rjd}(k)$$

Where $\rho_{rjd}(k)$ denotes the portion of density moving from region r to region d through the neighboring region $j \in J_r$.

With these elements it is possible to define the *intended outflow* of each region, defined through the fundamental equations of the triangular NFD, based on the function $q_r(\rho_r(k))$, that represents the intended outflow and is measured in vehicles/hour. The NFD is mathematically expressed as follows:

$$q_r(\rho_r(k)) = \begin{cases} \frac{q_r^C}{\rho_r^C}\rho_r(k), & if 0 \le \rho_r(k) \le \rho_r^C \\ w_r\left(\rho_r^J - \rho_r(k)\right), & otherwise \end{cases}$$

Where q_r^C indicates the capacity where the region operates at its maximum outflow is equal to $q_r^C = \rho_r^C u_r^f$, where u_r^f indicates the *free-flow speed* (km/h) at the determined time step, and ρ_r^C is the so-called *critical density*, i.e. the density when the intendent outflow equals the maximum outflow. Furthermore, ρ_r^J indicates the *jam density*, i.e. the density when both speed and outflow equal 0, and $w_r = q_r^C/(\rho_r^J - \rho_r^C)$ is the *backward congestion propagation speed*.

Similar to the intended outflow, two variables can be defined: $q_{rd}(k)$ and $q_{rjd}(k)$, indicating the *intended transfer flow* from region r to region d and the corresponding flow that passes through the neighboring region j, respectively.

$$q_{rd}(k) = \sum_{j \in J_r} q_{rjd}(k)$$

$$q_{rjd}(k) = \frac{\rho_{rjd}(k)}{\rho_r(k)} q_r(\rho_r(k))$$

The word *intended* in these variables indicates that the observed densities and quantities would transfer from a certain region towards another one and passing from the neighboring regions if there were not any capacity restrictions, but the intended transfer flow from a region r to a region j is actually restricted by the *inter-boundary capacity* $C_{rj}(\rho_j(k))$, that is the maximum flow that can pass through two neighboring regions, due to physical limits [111]. Therefore, the *actual transfer flow* can be defined as:

$$\tilde{q}_{rjd}(k) = \min\left(q_{rjd}(k), C_{rj}(\rho_j(k)) \frac{q_{rjd}(k)}{\sum_{y \in D} q_{rjy}(k)}\right)$$

Finally, taking all these factors into account, the dynamics of the vehicles transferring from the region r to region d can be described by the following equation:

$$\rho_{rd}(k+1) = \rho_{rd}(k) + \frac{1}{L_r} \tilde{d}_{rd}(\mathrm{k}) + \frac{T_s}{L_r} \sum_{j \in J_r} (\tilde{q}_{jrd}(k) - \tilde{q}_{rjd}(k))$$

Where T_s (min) denotes the simulation time step that regulates the evolution of the regional dynamics and L_r (km) is the total length of the roads in the region.

In words, the density of vehicles moving towards a certain destination from a selected region at a fixed time step is determined by the same density at the previous time step, plus the actual external demand and the difference between the quantity of vehicles entering the region from the neighboring regions to head towards the same destination and the quantity of vehicles leaving the region to pass through the neighboring regions.

This relatively simple NFD-based model has been used to create a vast multilayer simulation tool for the traffic dynamics in Manhattan, New York City, which is described in section 7.2.

Chapter 6

Spatial Analytics: applications

As explained in chapter 4, spatial enablement, i.e. the ability to add spatial information to data gathered in a certain spatiotemporal timeframe and to data analysis and visualization methodologies, has been a central concept throughout the design and the development of the PULSE project.

In this chapter, the main applications of this concept and of the analysis algorithms presented in chapter 4 are reported. In particular, after a brief presentation of the data integrated in the WebGIS, three works related to the PULSE project that apply spatial analytics to create new ways of analyzing public health and design interventions will be presented. These projects have been developed as part of the PhD work presented in this thesis and inserted in the PULSE framework. Their aim is to experiment new ways to apply spatial enablement to provide instruments to treat urban public health problems with the highest possible spatial resolution.

6.1. Data Integration in the WebGIS

The most visible integration of spatially enabled data has been performed in the WebGIS, that can be considered the best example of spatial enablement for data visualization, even if it features some interactive tools that are explained in chapter 3. As explained in section 3.2.2, the WebGIS collects a lot of data coming from different sources and of different kinds, namely health-related, demographic, socioeconomic, climatic and satellite images.

Most of the tabular data were already available in GIS-specific shapefile format (with the city usually subdivided in polygons representing different neighborhoods) with a clear spatial reference that allowed them to be overlapped to the other data. Other data were not already available in GIS format, but they contained a reference to the geographic position of the place where they were collected that allowed to add an identifier to make them representable in the polygons the city were divided into.

This has led to the necessity of a hard effort for the integration of heterogeneous data accompanied by several difficulties related to the spatial enabled features. Satellite images and raster images were already georeferenced, even if sometimes with different standards, and sensors data were usually provided with a geographic reference of the sensors' location, making it easy to integrate them inside the GIS maps. However, the integration of open data has been more complicated due to the lack of standards and harmonization procedures in the open data that can be found online. With the advancements of the current data collection technologies, there has been an increase in the capability to quickly collect and analyze large quantities of data, but unfortunately this is happening without a proper regularization and awareness from the public health authorities, leading to an increasing quantity of unorganized and chaotic data, difficult to import and integrate without a thorough supervision [113].

This lack of protocols and order hits data integration at different levels, as some effects are visible in the fact that different cities store and represent data with different standards, making it difficult to integrate them in a system that incapsulates more than one city such as the PULSE WebGIS, but even within the same city there are often problems to solve. For example, within PULSE the data integration had many unexpected delays and difficulties related basically to the following issues:

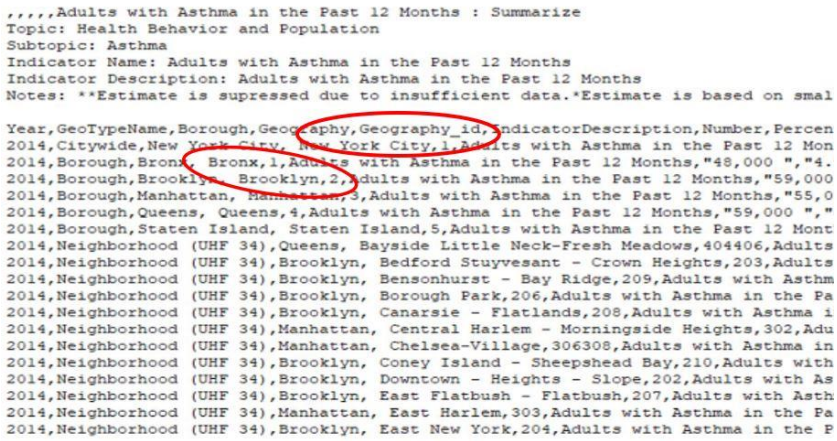
- Conventional spatial subdivisions: this problem has been encountered mainly for the city of New York (NYC), but the issue is not specific for this city. NYC has a very large amount of open data available, creating the potential for a lot of urban public health studies, but data are collected with reference to several different spatial subdivisions. The main ones are:

  - Boroughs: 5 polygons correspondent to the main districts of the city (Bronx, Brooklyn, Queens, Manhattan, Staten Island);
  - UHF 34 and UHF 42: two subdivisions correspondent to hospital districts (United Hospital Fund), with 34 and 42 polygons respectively;
  - CD 55, CD 59 and CD 71, where CD stands for *Community Districts*: three different subdivisions with 55, 59 and 71 polygons;
  - PUMA (Public Use Microdata Areas): 55 polygons different from the CD 55 subdivision;
  - NTA (Neighborhood Tabulation Areas): 195 polygons;
  - ZIP codes: 262 polygons.

  None of these subdivisions has vertices or edges that can be easily overlapped to vertices or edges of other subdivisions, so two or more phenomena that are measured with different spatial descriptions cannot be easily integrated without the application of extra spatial analytics.
- Geometric inconsistency: in several cases, even when the information was already available in GIS shapefiles, contained inconsistencies that required longer processing times. For example, areas that in one shapefile where indicated as identified polygons, in other shapefiles with the same spatial subdivision they were marked as holes with no data (Figure 6.1). Offsets in the shape of the polygons of different shapefiles were noticed as well, requiring hand work to fix them.

Figure 6.1: example of geometric inconsistency. Even if they are supposed to represent the same spatial subdivision, he polygons of the red layer do not match the polygons of the blue layer, that contains several islands that are inexistent in the red layer.

- Tabular data problems: even tabular data is not impervious to these kinds of problems, as numerous issues were encountered also in the integration of tabular data into the shapefiles of the GIS. Examples of this are the ambiguity of IDs and codes in the table and in the use of separators. Figure 6.2 shows an example of this related to the city of New York, where the code "1" is used both as a reference of the whole city and of the borough of the Bronx, and the comma is simultaneously used as the thousand separator and a field separator.



```
,,,,,Adults with Asthma in the Past 12 Months : Summarize
Topic: Health Behavior and Population
Subtopic: Asthma
Indicator Name: Adults with Asthma in the Past 12 Months
Indicator Description: Adults with Asthma in the Past 12 Months
Notes: **Estimate is supressed due to insufficient data.*Estimate is based on smal

Year,GeoTypeName,Borough,Geography,Geography_id,IndicatorDescription,Number,Percen
2014,Citywide,New York City, New York City,1,Adults with Asthma in the Past 12 Mon
2014,Borough,Bronx, Bronx,1,Adults with Asthma in the Past 12 Months,"48,000 ","4.
2014,Borough,Brooklyn, Brooklyn,2,Adults with Asthma in the Past 12 Months,"59,000
2014,Borough,Manhattan, Manhattan,3,Adults with Asthma in the Past 12 Months,"55,0
2014,Borough,Queens, Queens,4,Adults with Asthma in the Past 12 Months,"59,000 ","
2014,Borough,Staten Island, Staten Island,5,Adults with Asthma in the Past 12 Mont
2014,Neighborhood (UHF 34),Queens, Bayside Little Neck-Fresh Meadows,404406,Adults
2014,Neighborhood (UHF 34),Brooklyn, Bedford Stuyvesant - Crown Heights,203,Adults
2014,Neighborhood (UHF 34),Brooklyn, Bensonhurst - Bay Ridge,209,Adults with Asthm
2014,Neighborhood (UHF 34),Brooklyn, Borough Park,206,Adults with Asthma in the Pa
2014,Neighborhood (UHF 34),Brooklyn, Canarsie - Flatlands,208,Adults with Asthma i
2014,Neighborhood (UHF 34),Manhattan, Central Harlem - Morningside Heights,302,Adu
2014,Neighborhood (UHF 34),Manhattan, Chelsea-Village,306308,Adults with Asthma in
2014,Neighborhood (UHF 34),Brooklyn, Coney Island - Sheepshead Bay,210,Adults with
2014,Neighborhood (UHF 34),Brooklyn, Downtown - Heights - Slope,202,Adults with As
2014,Neighborhood (UHF 34),Brooklyn, East Flatbush - Flatbush,207,Adults with Asth
2014,Neighborhood (UHF 34),Manhattan, East Harlem,303,Adults with Asthma in the Pa
2014,Neighborhood (UHF 34),Brooklyn, East New York,204,Adults with Asthma in the P
```

Figure 6.2: example of tabular problem in a csv file relative to NYC data.

This work enlightened the necessity to increase awareness of the importance of spatial enablement from the data collection process, in order to create standards that ease the elaboration and integration of spatial information. In spite of this difficulties, a large wealth of data was integrated in the WebGIS, as described in section 3.2.2.

6.2. Spatial Enablement to study asthma hospitalizations in New York City

Thanks to the data integration performed in the PULSE system and to the exploration of the spatial analytic methods explained in chapter 4, it was possible to investigate more deeply the link between the exposome and health-related outcomes. In particular, this section reports a study concerning the link between environmental, socioeconomic and demographic factors and the hospitalization rate for asthma in New York City (NYC) [59].

Asthma is known to be a multifactorial disease, whose manifestation and exacerbation is related to a combination of genetic, social and environmental factors, Some of which are known to the scientific community [114], for instance the relation between air pollution and asthma complications [115] or the connection between asthma prevalence and demographic or socioeconomics conditions such as race, education, sex, and income [116]. In spite of this, there is a general lack of studies that address the problem at an intra-city level. Clustering neighborhoods and population according to their risk level could allow policy makers to better target their interventions.

A few studies focusing specifically on urban areas have been conducted, but they usually consider a coarse spatial subdivision which often corresponds to the whole city [117] or macroscopic grid-like subdivisions in the order of 10 × 10 km [118]. This is often a consequence of the scarce availability of well-integrated data regarding health and environmental exposure at a sufficient level of granularity to enable meaningful statistical analyses [59].

The study reported in this section shows an initial exploration of some of the high spatial resolution methods presented in chapter 4 and used to analyze the large heterogeneous data gathered in PULSE, demonstrating their necessity and usefulness. The study was performed on the urban area of NYC and its results confirmed that the hospitalization rate is related to a number of environmental and socioeconomic factors whose level of influence changes in the different areas of the city, enlightening, among other things, the importance of allying high-resolution spatial methods to study this sort of public health phenomena.

6.2.1. The asthma issue in New York

Asthma is one of the most diffuse respiratory problems in the world, and its prevalence, as shown in chapter 2, does not appear to be declining in most places. For this reason, asthma was chosen as one of the main target diseases in PULSE, as its incidence appears to be particularly high in urban areas. Among the test sites, New York has been facing an important asthma problem for many years. According to the Center of Disease Control (CDC), about 10% of all adults in New York City are asthmatic, with this prevalence rising to 17% in some corners of the South Bronx [119]. This percentage is higher than the 9.3% of the rest of the State, which is also higher than the nationwide prevalence of the whole United States, which is about 7.5%. The healthcare system of the State of New York spends 1.3 Billion dollars per year for asthma (second highest in the US).

The problem appears to be particularly present in some neighborhoods in the Bronx.

According to the community health profiles [120] published by the New York City Department of Health in 2015, 0.3% of children aged 5 to 14 were hospitalized for asthma in New York City in 2013, 0.7% were hospitalized in the Bronx, and 1.2% in the South Bronx neighborhoods of Mott Haven and Melrose. In the same year, the percentages of adults hospitalized were 0.2% globally in New York City, 0.5% in the Bronx, and 0.7% Mott Haven and Melrose. According to public schools records used by the NYC Department of Health, the percentage of asthmatic children ages 5-14 years has increased from 2.4% to 3.5% from 2010 to 2014, and asthma appears to hit harder the African American and Latino populations, for reasons that are still partially unknown.

6.2.2. Data Sources and Pre-processing

Several sources of data have been used to carry out the analyses reported in this section. Some of the data were gathered from open repositories (NYC is one the cities with the largest availability of recent data), other were provided to the PULSE consortium by the New York Academy of Medicine. The data to be collected were chosen based on the elements that appeared to be most related to asthma exacerbation according to previous evidence in literature.

Several variables, listed and described in Table 6.1, have been chosen for this study. For air pollution, we chose the PM2.5 and ozone concentrations, as among all the pollutants they appear to be particularly relevant in asthma exacerbation [17], [121] and open data about their monitoring in NYC is easily available. We then chose a set of other environmental data related to asthma (e.g., industrial land use) and some socioeconomic factors that appeared to be relevant in previous research [115], [118, p.].

Table 6.1: description of the data used in our study with the correspondent data sources.

Type	Description	Source	Year	Sample Size
Health-related	Hospitalization rate: number of people hospitalized for asthma over total population	SPARCS, NYC Data Portal	2014	42 observations (one for each UHF42)
Environmental	PM2.5 yearly average	NYC Data Portal	2014	42 observations (one for each UHF42)
Environmental	Ozone summer average (from June to September)	NYC Data Portal	2014	42 observations (one for each UHF42)
Environmental	Percentage of land used for industrial activities	Data2go.nyc	2017	59 observations (one for each CD59) – 42 after interpolation
Environmental	Recycling rate	Data2go.nyc	2010	59 observations (one for each CD59) – 42 after interpolation
Demographic	Age: percentage of population aged <18, average age at hospitalization and percentage of population aged >65	SPARCS	2014	174 observations (one for each zip code) for average age – 42 (one for each UHF42) after interpolation. 42 observations for the percentages
Demographic	Race: percentage of people identifying as Black, Hispanic, Asian, White, Other/Unknown	NTA	2014	195 observations (one for each NTA) – 42 (one for each UHF42) after interpolation
Socioeconomic	Poverty rate	NTA	2014	195 observations (one for each NTA) – 42 (one for each UHF42) after interpolation
Socioeconomic	Medicaid coverage	NTA	2014	195 observations (one for each NTA) – 42 (one for each UHF42) after interpolation

The socioeconomic data used in this paper are freely available in the NYC Neighborhood Health Atlas website [122], the PM2.5 historical data have been downloaded from the NYC Environmental & Health Data Portal [123]. The data regarding the percentage of land used for industrial purposes, the obesity rate and the recycling rate were taken from the data2go.nyc website [52]. The information regarding age and race of hospitalized people has been acquired from the SPARCS [55] limited 2014 dataset, which is not freely available and has been provided by the New York Academy of Medicine.

Some preprocessing operations had to be carried out in order to uniform the spatial description of the data, as different datasets were described using a different spatial subdivision, among the ones presented in section 6.1. In particular, we decided to adopt the UHF42 subdivision to visualize the results, as most of the data considered is natively available for it. Since the different subdivisions used in NYC do not share vertices and edges if overlapped, we applied a simple data harmonization algorithm as follows: let's consider a polygonal subdivision for which a certain variant is available, for each polygon P_i, the variant value v_i is known. If we consider another polygon P_0, belonging a different subdivision, in general it won't coincide with any P_i and, instead, will overlap to several of them. The estimated v_0 can be obtained by the weighted sum:

$$v_0 = \frac{\sum_i v_i A_i}{\sum_i A_i}$$

Where A_i is the area of intersection between P_i and P_0 if the intersection is non-empty. In words, the value of a phenomenon or indicator in any spatial subdivision is reported in the UHF42 subdivision constructing the polygons as the sum of the values in the polygons of the other subdivisions, weighted for the overlapping areas.

6.2.3. Spatial Clustering analysis

The first step of our exploratory analyses consisted in a set of spatially-enabled clustering analyses. Firstly, we tried to assess the relationship between air pollution and asthma hospitalizations using the spatial clustering method explained in section 4.3.1, considering the rate of asthma hospitalizations and the yearly PM2.5 concentration as features. We performed the clustering multiple times, each one considering the data of a specific year from 2012 to 2014, obtaining systematically the same clusters as those shown in Figure 6.3a, showing 2014 data. Some interesting phenomena can be noticed in the clustering results:

- In 7 out of 10 clusters, low PM2.5 concentrations correspond to low asthma hospitalization rates.
- The remaining 3 clusters, all correspondent to areas located in the borough of Manhattan, are in contrast with the rest of the clusters, as they show high pollution levels and low hospitalization rates;
- The Bronx, East Harlem and some neighborhoods in North/Central Brooklyn crossed by important highways (e.g., Brooklyn-Queens Expressway, Long Island Expressway) have the highest hospitalization rates of and the highest pollution levels of all the neighborhoods excluding the wealthiest areas of Manhattan.

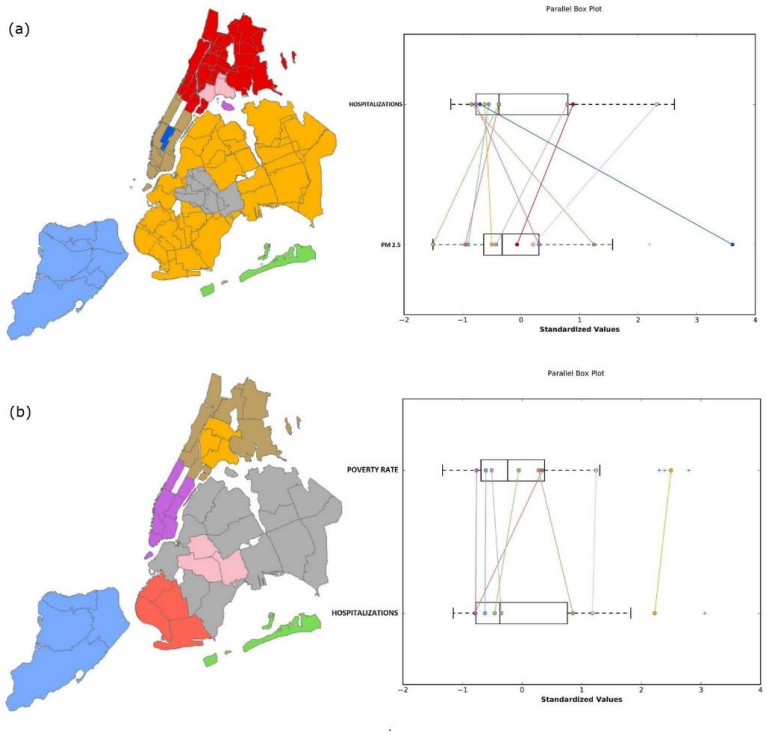


Figure 6.3: Results of the spatial clustering considering **(a)** the average 2014 PM2.5 concentration and the asthma hospitalization rate and **(b)** the poverty rate. On the left, a color-coded map of the clusters and on the right a parallel box plot that shows the relation between the parameters' distributions.

These results suggest that although there is clear and documented relation between air pollution and asthma, also other factors, probably socio-economic, can play an important role. This is demonstrated by the Manhattan case, as in our results Manhattan, known to be the wealthiest borough of the city, has for the most the highest pollution concentrations and the lowest hospitalization rates. To further explore this point, we performed another spatial clustering operation in which we focused on the link between child asthma hospitalizations and poverty. Figure 6.3b presents the results from the spatial clustering analysis using asthma hospitalizations and poverty rate. Apart from some neighborhoods in West Brooklyn, there seems to be an association between high

poverty rates and high hospitalization rates, as in the same areas where the hospitalization rate is high (i.e., south Bronx and north/central Brooklyn), the poverty rate also appears to be higher than the rest of the city.

6.2.4. Geographically Weighted Regression to unveil relations between asthma hospitalizations and environmental and socioeconomic factors.

Given the initial insights shown by the spatial clustering results, we performed a more thorough analysis to better assess the relations between asthma hospitalizations and many other factors, not only related to pollution.

First of all, we performed a univariate standard linear regression using 2014 asthma hospitalizations as dependent variable and the average yearly PM2.5 concentration for the same year as covariate. Our results confirm the same phenomena pointed out by the spatial clustering: PM levels do not have a significant impact on the hospitalization rate (*P*-value 0.985, Correlation Coefficient 0.003). Considering the spatial clustering results, in which it's clear that in midtown Manhattan the relationship between air quality and hospitalizations is different than the rest of the city, we carried out the same linear regression excluding all the neighborhoods of Manhattan but East Harlem, Central Harlem and Washington Heights (since they don't share the same contrast between high pollution and low hospitalization rate), and we obtained a significant relationship between air pollution and hospitalizations with a moderate positive correlation (*P*-Value 9.62×10^{-4}, Correlation Coefficient 0.534).

We then repeated the analysis using the 2014 poverty rate as covariate to predict the asthma hospitalizations, results show a very strong correlation (*P*-Value 5.93×10^{-13}, Correlation Coefficient 0.855). From these results it could be assumed that the average yearly value of PM2.5 is not a good measure to study the effects of air quality on asthma, as local and brief but potentially dangerous peaks are not visible. Furthermore, these results confirm that, even if air quality plays a role in determining the number of asthma hospitalizations in several areas of the city, it appears that socioeconomic factors are much more decisive when all the city is

considered. Hence, the spatial dimension cannot be neglected in studying the hospitalization rates in a city as NYC, in which there are pronounced socioeconomic and environmental differences among the neighborhoods. For these reasons, further analyses employing the Geographically Weighted Regression (GWR) method explained in section 4.2.2 were performed in order to correct for the local effect of socioeconomic factors in the different areas of the city. We applied our algorithm several times using a regular spaced grid of points distant 1 km from each other and setting 5 km as threshold, and tested different covariates singularly, specifically: PM2.5 and ozone concentration in the same year (2014), poverty rate, percentage of the population identifying as Black, obesity rate, percentage of population aged under 18 or over 65, recycling rate. After these tests, we also created a multivariate model to explore the usefulness of this method also in the analysis of multiple covariates. Figure 6.4 shows how the GWR was applied to obtain a set of spatially enabled regressions, one for each dot of the gird overlapped on the city GIS maps, starting from data represented in the UHF42 subdivision.

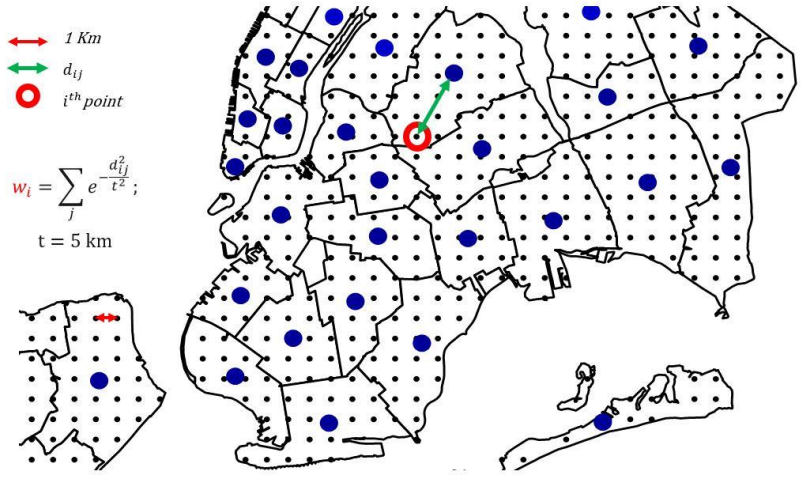


Figure 6.4: graphical representation of how GWR works to compute a linear regression for every dot overlapped to the GIS map of the city. The weights are calculated based on the distance between each dot and the centroids of the polygons, where the measured values are conventionally located.

6.2.4.1. Air Pollution

According to our Spatial Clustering results air pollution, in particular PM2.5 concentration, doesn't seem to have a significative impact on the asthma hospitalization rate in the whole city. Our results from the GWR confirm this hypothesis, since R^2 is low in most of the city, and in Manhattan β_1 is even negative, indicating the contrast between the high pollution level and the low hospitalization rate. To investigate also other pollutants, we applied the GWR also to the average concentration of Ozone, selecting as dependent variable a subset of hospitalizations occurred from June to September 2014, since the open data about ozone available is only referred to summer months. Results are not reliable, since in most of the city R^2 is low, and β_1 is generally negative. The overall correlation is −0.2186 and the *P*-Value is 0.164. This could be due to (i) higher influence of socioeconomic conditions than air pollution in the hospitalizations rate, (ii) low significance of averaged pollution data of a long period, that hides peaks and daily variations that can have an important effect on asthma. The GWR results regarding this part are shown in Figure 6.5.

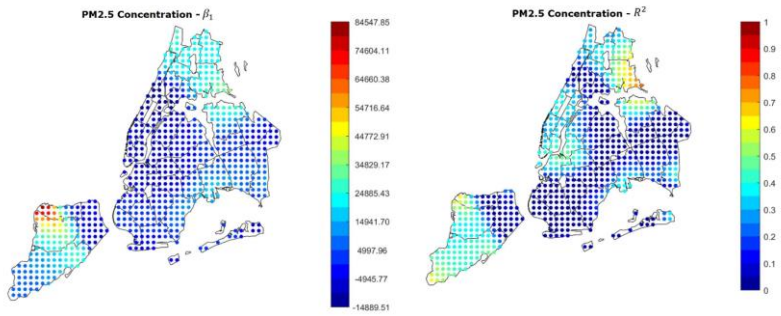


Figure 6.5: Results of the GWR using the average yearly PM2.5 concentration as covariate. The yearly value is not a good predictor in most of the city.

6.2.4.2. Race

Previous research [124] suggested that race and ethnicity can be related to asthma risk and development. Table 6.2 shows the results of a first explorative analysis on the hospitalization rates per race in the five boroughs: each value is calculated as the number of hospitalizations divided by the number of people identifying as the specific race considered. This table shows that in the Bronx, Brooklyn and Queens, where the overall asthma rate is the highest, the hospitalization rate is higher for Black people. Previous studies [125], [126] suggested that in several areas of the USA Black people and Latinos are more easily exposed to damaging pollutants since they usually live in areas close to industrial facilities or large highways. Figure 6.6 shows that the higher concentration of Black people in NYC is in the same areas where the higher hospitalization rates are. This phenomenon is mostly confirmed by the results of the GWR applied with the percentage of Black people as covariate, the results of which are also visible in Figure 6.6 (c,d). The overall P-Value is 4.94×10^{-4} and the correlation is 0.5143.

Table 6.2: percentage of people belonging to each race group that were hospitalized in 2014 in each borough.

Borough	Black	Hispanic	White	Asian	Other
Bronx	0.0213	0.0015	0.0011	0.0008	0.031
Brooklyn	0.5951	0.1520	0.5606	0.1043	0.6203
Manhattan	0.0149	0.01	0.0144	0.0017	0.0515
Queens	1.3909	0.3003	0.6689	0.0287	0.4692
Staten Island	0.069	0.006	0.0026	0.001	0.0211

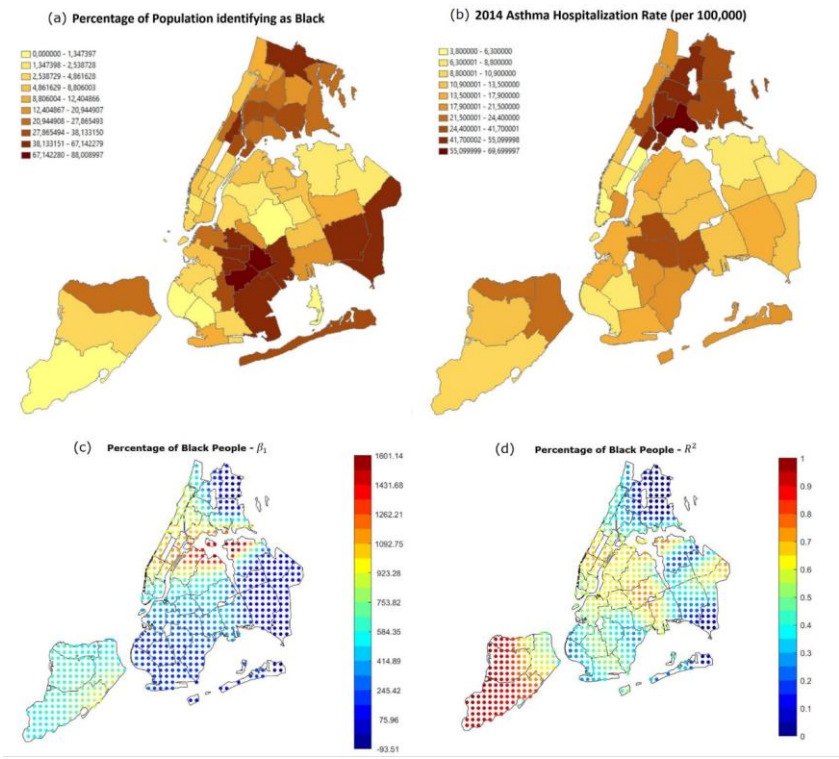


Figure 6.6: **(a)** Percentage of black people in the 42 districts in NYC. The highest numbers can be seen in the Bronx, in Harlem and East Harlem, South-East Queens and East Brooklyn, especially Crown Heights, Flatbush and Brownsville. Most of these areas are the same in which the hospitalization rate is high, visible in **(b)**. **(c)(d)** GWR results using the percentage of people identifying as black as covariate. The correlation is positive and reliable in most of the city, especially all Manhattan, the Brooklyn-Queens border, Staten Island and South Bronx.

6.2.4.3. Poverty Rate

The Spatial Clustering analysis suggested a relation between poverty and asthma in the city, confirmed by the a-spatial linear regression. To investigate if and how this correlation varies throughout the city, we applied GWR also to the 2014 poverty rate in all the neighborhoods. Figure 6.7 shows the β_1 and R^2 parameters in the different areas of the city. It is noticeable that in most of the city β_1 is positive and R^2 has values > 0.5. This means that in most

of the city the probability of observing asthma attacks increases with the poverty rate. In contrast with this, an area between south-west Brooklyn and east Staten Island shows values of R^2 close to 0, meaning that in those neighborhoods the relation found is not reliable, and further analyses on those neighborhoods are required.

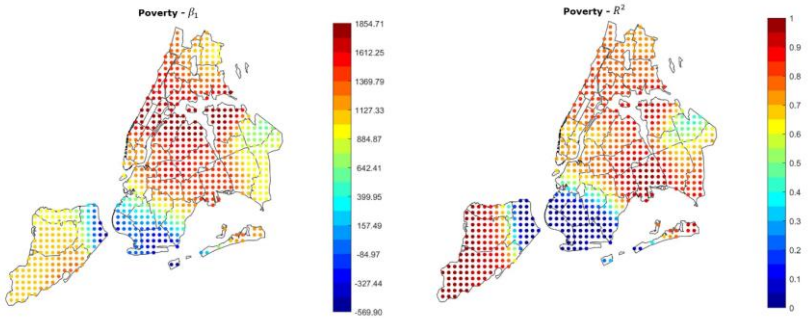


Figure 6.7: GWR results using the poverty rate as covariate. The relation is strongly positive and reliable in most of the city, the only exception is an area between South-west Brooklyn and East Staten Island.

6.2.4.4. The effect of age

It has been previously demonstrated that children and teenagers are more prone to asthma than adults [127]. Following the uneven distribution of services and cost of living, the population age distribution changes among the different neighborhoods in NYC: as shown in Figure 6.8 and 6.9, population aged under 18 tends to concentrate in the central areas of the Bronx and east Brooklyn/west Queens, whereas the highest rates of population aged more than 65 can be found in Manhattan, east Queens, peripheral areas of the Bronx and south-west Brooklyn. Furthermore, we found that the age distribution of patients hospitalized for asthma in 2014 has the same shape in all neighborhoods, with a bimodal shape that shows a primary peak in young age and a secondary peak after the age of 40 [59]. Running a one-way ANOVA test, we found that age at hospitalization is significantly lower in the Bronx than in all the other boroughs, as it is lower in Brooklyn and Queens if compared to Manhattan and

Staten Island (the average age data are reported in Table 6.3). This is an interesting finding considering that the age-adjusted asthma prevalence per 100 individuals is 6.2 in the Bronx, 3.8 in Brooklyn, 4.6 in Manhattan, 3.7 in Queens and 5.7 in Staten Island [128] (data of the year 2002), therefore the areas with the higher prevalence are not the same with the higher hospitalization rates.

Table 6.3: age at the moment of admission to the hospital for asthma in each borough.

Borough	Average Age	Standard Deviation	Observations
Bronx	30.85	28.24	4289
Brooklyn	39.88	30.14	3927
Manhattan	44.92	29.39	1832
Queens	41.45	32.22	1954
Staten Island	48.55	24.92	397
All	38.03	30.09	12399

Figure 6.8 (c,d), shows that GWR applied to percentage of people aged under 18 (Correlation 0.6389, *P*-Value 5.27×10^{-6}) shows that in the areas with the higher prevalence and hospitalizations rate (i.e., the Bronx and east Brooklyn/west Queens), β_1 is positive and R^2 is high, this happens also in central Staten Island; using the percentage of population over 65 as covariate (Correlation -0.5342, *P*-Value 2.69×10^{-4}) it can be noticed that in the same areas β_1 is negative, therefore high rates of older people prevent hospitalization rates from rising (Figure 6.9).

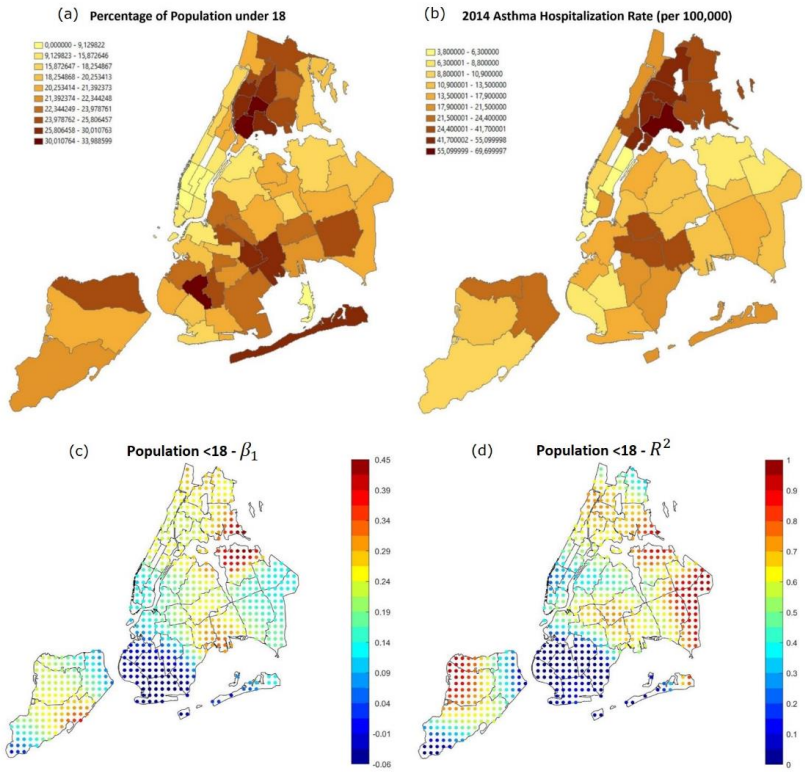


Figure 6.8: **(a)** Percentage of population aged under 18, most of it is concentrated in the Bronx and central-east Brooklyn, where also the asthma hospitalization rate is higher, as visible in **(b)**. **(c) (d)** Results of GWR using the percentage of population aged under 18 as covariate. The correlation is positive and reliable in most of the areas with higher hospitalization rate.

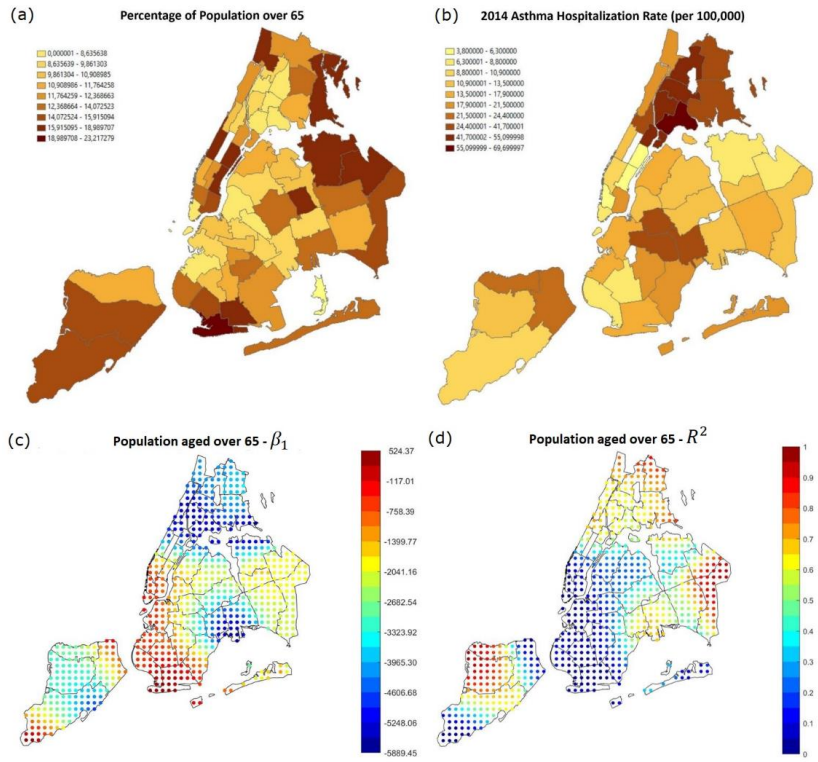


Figure 6.9: **(a)** Percentage of population aged over 65, most of it is concentrated in Manhattan, Staten island, East Queens and South-west Brooklyn, where the asthma hospitalization rate is lower, as visible in **(b)**. **(c) (d)** Results of GWR using the percentage of population aged over 65 as covariate. The correlation is negative and reliable in most of the areas with low hospitalization rate.

6.2.4.5. Other socioeconomic variables

A lot of other socioeconomic variables were tested using GWR, most of which showed interesting results. One of this was health insurance coverage, as several studies demonstrated that people with no insurance or with public insurances such as Medicaid and Medicare tend to visit the ED more often that people with private insurance [129], [130]. Our analysis confirmed that even insurance coverage can be a predictor for asthma hospitalizations in the high-rate areas of NYC: with a global correlation coefficient of

0.7138 (*P*-Value 11.1 × 10^{-7}), our results showed that the areas with the higher hospitalizations rate are the same with the higher Medicaid coverage, and also those with the higher β_1 and R^2 [59].

Another factor that has been identified as related to asthma in previous studies is obesity [131]. Figure 6.10 shows that the higher obesity rates in NYC are between the Bronx, Upper Manhattan and East Harlem, south-east Brooklyn and north Staten Island. According to the GWR results, also visible in Figure 6.10, there is a positive relation between obesity and hospitalizations in all the city, with high significance value in Upper Manhattan (especially East Harlem), Queens/east Brooklyn and north-west Staten Island. The overall correlation coefficient is 0.679 and the F-Statistic *P*-Value is 7.69 × 10^{-7}.

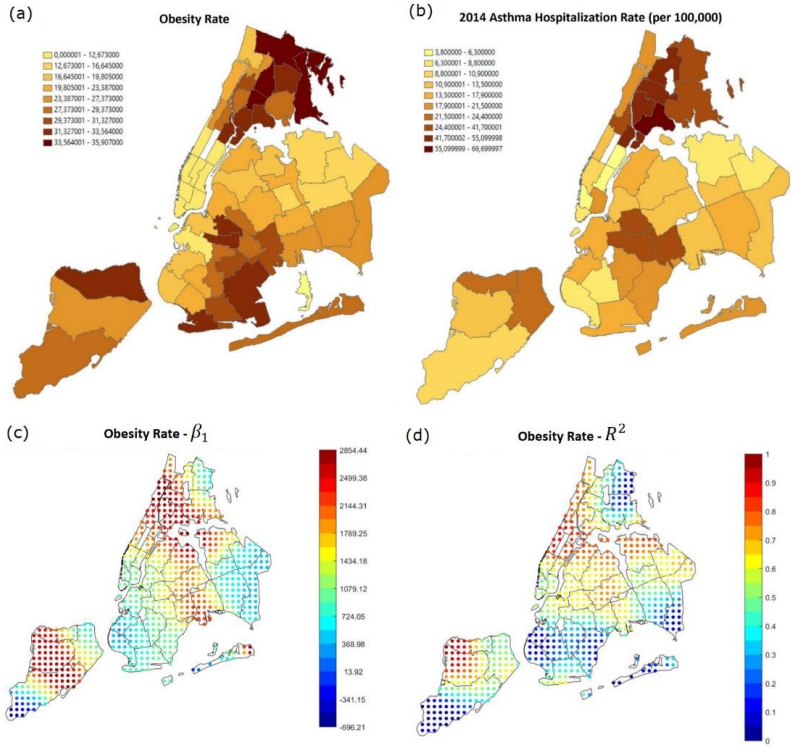


Figure 6.10: **(a)** Obesity rate in NYC, it is noticeable that the higher rates are in the same areas in which the asthma hospitalization rate is higher **(b)**. **(c)(d)** GWR results using obesity as a covariate. The relation is generally positive.

Another environmental factor that appears to be related to asthma is the percentage of land used for industrial activities, as our results, shown in Figure 6.11, reveal that there is a positive significant relation in the Bronx, where the hospitalization rate is higher, and in some other spots in Brooklyn and Queens. This could lead to the assumption that even if in a long-term measurement air pollution is higher in other neighborhoods, like in Manhattan, the presence of a lot of industrial sites could provoke brief local pollution peaks that could be a threat for people with asthma, this topic requires further investigation.

Also the garbage recycling rate appears to have an influence, as in most of the city there is a moderate negative relation that

indicates that a high recycling rate corresponds to lower asthma hospitalization occurrences.

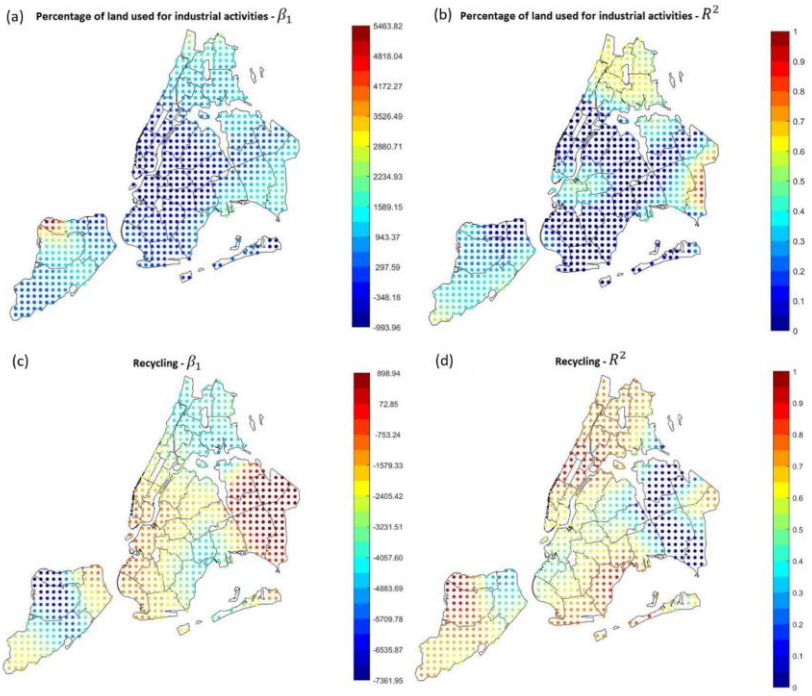


Figure 6.11: **(a)(b)** GWR results using the percentage of land used for industrial or commercial activities as covariate. The relation is quite strong and positive in the Bronx. **(c)(d)** GWR results using the percentage of recycled garbage as covariate. Excluding some areas in East Queens, the relation is quite strong and negative throughout the city.

6.2.4.6. Multivariate Analysis

GWR can be also performed testing several covariates at the same time, creating a multivariate model. Since the univariate results, as well as several studies published in literature, highlighted how poverty rate and race play an important role in increasing the probability to get hospitalized for asthma, we created an example of multivariate GWR that combines poverty rate and percentage of people identifying as Black and Hispanic.

The underlying model can be described by the following equation, valid for each point where the GWR is performed:

$$Hosp = \beta_0 + \beta_1 \cdot Poverty + \beta_2 \cdot \%Black + \beta_3 \cdot \%Hispanic$$

Where *Hosp* represents the hospitalization rate. Results are visible in Figure 6.12. On the left side of the image, maps of the β coefficients are shown, whereas panels in the right side show the correspondent significance maps based on the t-statistic values. In detail, we created 3 significance levels: Non-Significant (NS), Partially Significant (PS), Significant (S). The correspondent t-statistic threshold values are 1.96 (5% confidence level) and 2.58 (1% confidence level). Figure 6.13 shows the percentage of Hispanic people in the different neighborhoods on the left side (useful for the analysis reported below in this section) and the global R^2 of the model. Several interesting phenomena can be noticed in these figures:

- R^2 is extremely high in all the region, therefore the linear model is globally reliable.
- Considering poverty and percentage of Black population, the correspondent β are always positive, indicating a positive correlation between either of these factors and the hospitalization rate.
- In general, the higher the β, the higher the level of significance. Therefore, in the neighborhoods in which we found that high variables' levels lead to high hospitalization rates, the found relations are significant.
- Low significance levels could be due to other confounding variables and to a smaller quantity of data available. For instance, Figure 6.12, left side, shows the values of the variable associated to β_3, i.e., percentage of Hispanic people. It can be noticed that most of the Hispanic population is concentrated in the Bronx, Upper Manhattan (Harlem, East Harlem and Washington Heights), in central and west Queens and some areas of east Brooklyn (Bushwick and south of Highland Park), plus some isolated spots in west Brooklyn (Sunset Park) and north Staten Island. Apart from these last isolated spots, in the same areas in

which the concentration is higher, also the significance of the correspondent beta is high. Hence lower significance corresponds to higher scarcity of data.

These results show that even multivariate geographical analysis can be helpful to describe and visualize important public health phenomena and discover the relations among different factors.

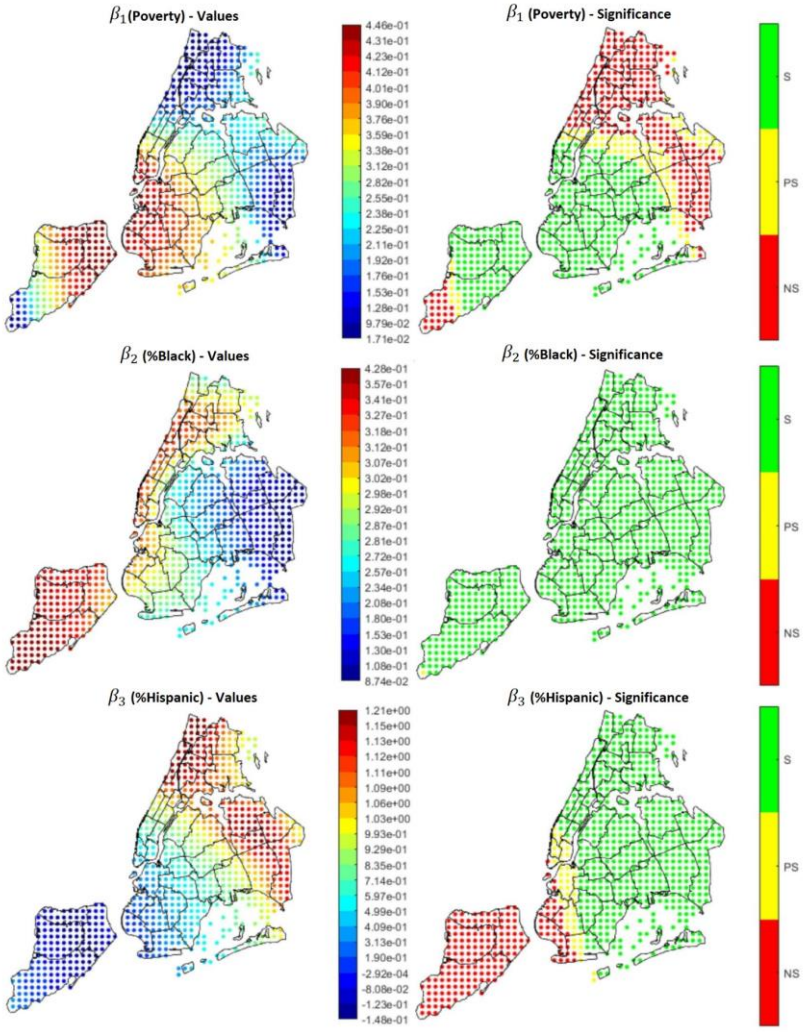


Figure 6.12: Coefficients (left side) and their significance level (right side) for the multivariate model

$$\text{Hosp} = \beta_0 + \beta_1 * \text{Poverty} + \beta_2 * \%\text{Black} + \beta_3 * \%\text{Hispanic}$$

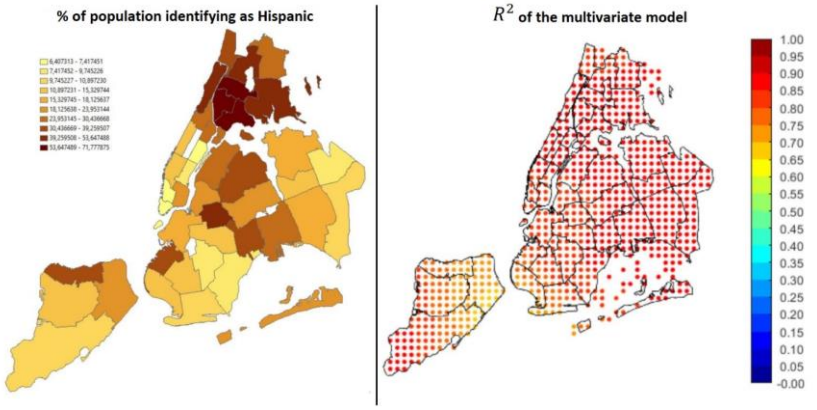


Figure 6.13: On the left: percentage of Hispanics over the total population. On the right: global R^2 of the model.

6.2.5. Main findings

Past research has already demonstrated the existence of a link between asthma hospitalizations and PM2.5 concentration [132], and also the relation between asthma and socioeconomic factors has already been investigated. Nevertheless, asthma remains an extremely complicated multifactorial disease, and past studies that have been conducted on urban areas usually focused on large geographical areas, if not the whole city. Studying the problem with a higher spatial resolution can help to inform policy makers and citizen themselves on the situation in the different neighborhoods, in order to organize targeted interventions and prevent new hospitalizations from happening as much as possible. The PULSE project was founded on this principle and aimed at spreading it to all the world's biggest cities.

The results of this study showed that socioeconomic factors have an extremely important impact on asthma hospitalization: a higher risk of asthma is associated with poverty, race/ethnicity, age (risk increases in younger patients), obesity, proximity to industrial areas, proximity to low recycling areas. We found that poor people and people without insurance or covered by Medicaid are more

likely to visit the hospital for asthma, in accordance with previous studies that showed a different use of the hospitals from people with different kinds of insurance, demonstrating that people without insurance or with public programs such as Medicaid are more likely to visit ERs. We then found that the age at hospitalizations is lower in the Bronx compared to the other areas of the city, most of the highest peaks of hospitalization rate are in this borough too, as well as a generally higher poverty rate. In a limited part of the city that includes south-West Brooklyn and East Staten Island, some factors such as poverty, obesity and insurance coverage have little or no effect on the hospitalizations rate. Deepening the analysis, we are able to hypothesize that this is due to particular conditions of the environment (i.e., pollution is low, there are no factories, garbage disposal is adequate) and of the population (medium income and generally adult) that prevent hospitalizations from happening. Further investigation in these areas is indeed required, nevertheless this demonstrates that spatial enablement is necessary to aid public health in big cities, providing useful tools both for visualization and discovery.

Concerning pollution, we analyzed the effects on asthma hospitalizations due to PM2.5 and Ozone, obtaining a bland relation in 4 out of 5 boroughs for PM2.5, and unreliable results for Ozone. It should be noted that further investigation on this topic might be required, as yearly averaged data could be unsuitable to estimate the real effect of pollution, since it could hide temporary peaks that could affect the health of asthmatic people causing short-term exposure effects. As explained in section 2.1.2, the lack of pollution data with high granularity in space and time is a common problem in a lot of cities, for example NYC has only 13 official monitoring stations on a 784 km^2 area.

From a practical point of view, it cannot be neglected that some of the factors that have been addressed in this study are manageable more easily than others by public health authorities, for example traffic reduction and car improvements laws can limit air pollution, delocalization of industrial activities from the city, combined with an increase of the number and the extension of green areas can improve the environment, food policies and sensibilization campaigns can reduce obesity etc. On the other hand, it is quite difficult for a local authority to intervene on factors such as poverty and insurance, that are related to central health policies and rules. Nevertheless, our findings demonstrate the

utility of our approach and provide an example of the importance of gathering data with a high spatial resolution and using highly spatially enabled techniques to address health problems in urban environments.

6.3. Transfer Learning for urban image clustering

As a tool studied to help health policy makers in the intervention planning, the PULSE dashboard incapsulates a set of tools that are meant to provide assistance in all the possible challenges of the process that starts with detecting the public health problems and culminates with designing a solution based on proper analyses ad critical evaluations. One of the problems that has to be addressed is efficiency in the study and application of urban planning strategies. This section reports a study performed in this contest, published on the journal *Sensors* [133], where we created an analysis pipeline based on a deep learning strategy that allows to cluster urban areas together and find similarities in the health of the inhabitants, creating a quick categorization of the city that can speed up the intervention design process.

6.3.1. The urban planning challenges

The study reported in section 6.2 highlights the necessity to study public health at a neighborhood level in a big city, given the high heterogeneity of the environment and the socioeconomic panorama. Following this idea, also public health interventions should be based on data with high spatial granularity and should address problems related to small areas, in order to contribute to the creation of a better environment in the whole city. Although spatial enablement, georeferencing of information and advanced spatial analytics facilitate the data gathering and analysis process, the intervention design process could still be a long and difficult process. In a big city with hundreds of neighborhoods, planning interventions on each neighborhood means repeating hundreds of times the same process, requiring long times and high costs to reach measurable results.

For this reasons, it can be of interest to healthcare planners and city decision-makers to have instruments able to find clusters of city areas that share similar urban structures and to analyze some behavioral indexes of their residents, in particular to see potential correlations and to plan similar interventions in the different clusters, even if such clusters refer to areas that are geographically far away. This would lead to a notable contraction of time and expenses. The research question that motivated the study reported in this section was whether there is a way to cluster neighborhoods together based on their urban landscape in a way that also their public health situation is similar.

Recent advances in machine learning and deep learning enable the design and implementation of novel data analysis pipelines that allow fusing heterogeneous data sources to extract novel insights and predictive patterns [134], [135]. These approaches seem particularly suitable to increase our insights in the relationships between the urban landscape of cities and the behavior of their residents, with particular focus on well-being and healthcare indexes.

The link between urban structure and health has already been investigated in the past, for instance Krefis et al. wrote a systematic review in 2018 [136], showing that the link between green areas and health have been addressed by several studies, but also pointing out a lack of interdisciplinary studies that approach the complexity of urban structure, its dynamics and links with wellbeing. Some socio-technical studies have been performed as well, such as the one conducted by Tavano Blessi et al. in Milan, Italy [137], in which they analyzed survey data to determine the influence of urban green areas in the precepted wellbeing of the population. In addition, deep learning has been already used in a number of studies on these topics, like for example the one by Helbich et al. [138], where they used deep learning on street view images to determine whether there was an association between the presence of green/blue areas and geriatric depression. Their results support this hypothesis, even though causal relationships were not fully investigated.

These studies are either very general (i.e., they focus on general wellbeing) or very specific, since they investigate the links between a specific environmental factor and a specific condition. In our study, we present an analysis pipeline to try to answer the same research questions, but investigating the influence of urban

structures in a series of preventable health complications, prevention strategies and behavioral health risk factors that mirror the association between the social structure of the city and the physical urban environment. To this end, we resorted to the capability of deep neural models to process images and to correlate them with outcome measurements, such as health indexes.

Deep neural models provide flexible instruments to perform the non-linear approximation of a variety of multivariate functions and to extract latent variables from a data set. In dependence of the nature of the input data set, different architectures can be exploited, ranging from the combination of many convolutional layers in the case of images, to the use of long-term/short-term networks in the case of time series and speech/text data.

Recently, an increasing number of papers are using deep learning to examine the relationships between the urban landscape and some environmental data, as well as with citizens' behavioral data [139]–[141].

One of the main limits of deep learning models is related to the large data sets needed to reliably estimate their parameters. In fact, in order to be able to gain the advantage of their capability of encoding even the finest details that can be important to map input data, large data sets are necessary in order to avoid getting trapped into noise resulting in overfitting and poor parameters estimates.

In order to deal with this problem, it is possible to resort to an increasing set of pre-trained deep learning models that can be used for the task of transfer learning [142], i.e., models that are able to represent the input space into a set of latent variables on the basis of a mapping mechanism, usually a deep neural network, learned on a large (external and independent) data set, so that the relationships between such latent variables and the outcomes can be later learned on a specific and smaller data set.

6.3.2. Data Sources

Our analysis is based on two data sources: NYC high resolution images and healthcare data coming from the 500 cities project (see chapter 3). NYC images have been collected by the "The National Agriculture Imagery Program" (NAIP) that acquires aerial imagery during the agricultural growing seasons in the continental United States. In particular, we have downloaded an image of NYC having

an original resolution of 0.5 m and have downsampled it to 2 m which allows to have a fine-grained representation of the aerial urban landscape (see Figure 6.14).

As it will be explained in the following, the reason for the downsampling is that the original large image has been subdivided into tiles and the neural network adopted can accept images having maximum size of 299 pixel; we had to tune the ground resolution in order to have meaningful tiles, embracing a sufficiently-sized area.

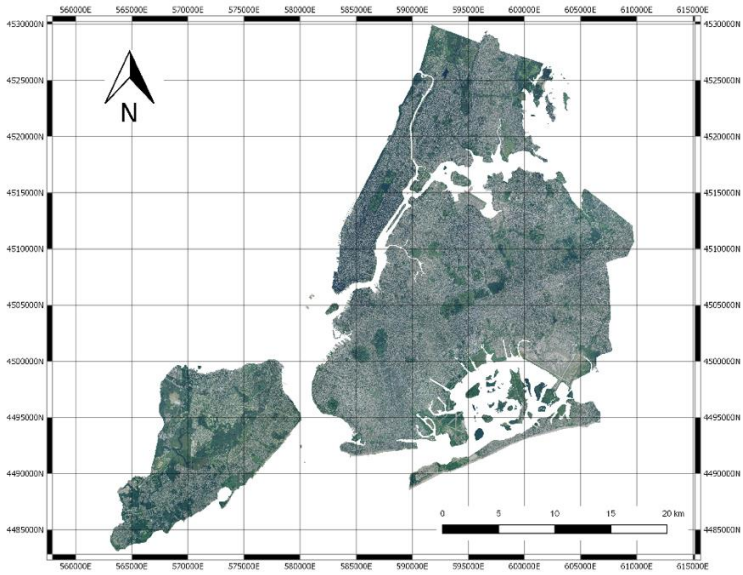


Figure 6.14: NAIP image of New York City.

Healthcare data have been extracted from the repository made available by the 500 Cities project, introduced in chapter 3. "500 cities" is a collaboration between CDC, the Robert Wood Johnson Foundation, and the CDC Foundation [1]. The project provides estimates for chronic disease risk factors (unhealthy behaviors), health outcomes, and clinical preventive service use for the largest

[1] https://www.cdc.gov/500cities/index.htm

500 cities in the United States. Such estimates are provided for each *census tract* of a city.

Each American state is divided into counties, and the area of each county is further organized into *census tracts*. *Census tracts* are conventional geographical entities within the US counties [143] defined for census operations, and they represent the smallest territorial entity for which population data is available [144].

The last census in the United States was organized in 2010 and the number of *census tracts* which the American territory is divided into is 74,134. Generally, each of them is different from the other with respect to several features such as the population (between 1,200 and 8,000 people) and the spatial dimension, which depends on the area density [145].

Figure 6.15 represents the city of New York divided into its *census tracts*, that, to be precise, are 2,166.

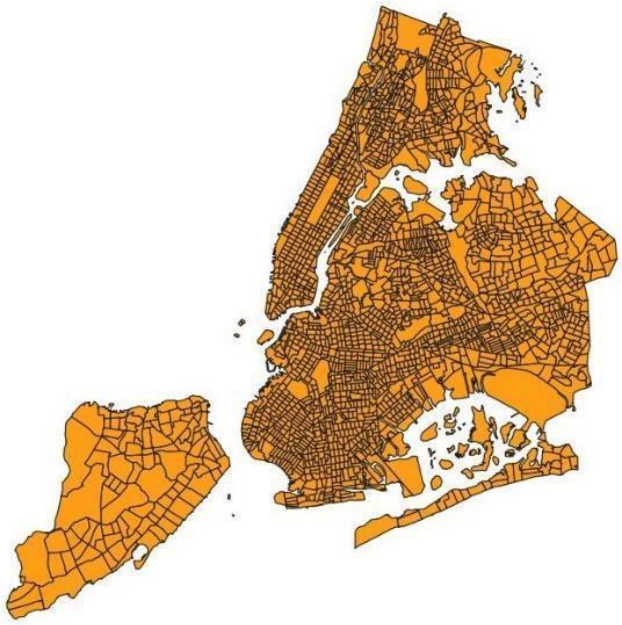


Figure 6.15: NYC subdivided into its census tracts.

The 24 chronic diseases measures provided by the project are listed in Table 6.4. They include major risk behaviors that lead to

illness, suffering, and early death related to chronic diseases and conditions, as well as the conditions and diseases that are the most common, costly, and preventable of all health problems.

Table 6.4: 500 Cities measures grouped by category. The 24 measurements include 13 health outcomes, 9 prevention practices and 5 unhealthy behaviors.

Category	Measure
Health outcomes	Current asthma among adults aged >=18 years High blood pressure among adults aged >=18 years Cancer among adults aged ≥18 years High cholesterol among adults aged >=18 years who have been screened in the past 5 years Chronic kidney disease among adults aged ≥18 years Chronic obstructive pulmonary disease among adults aged >=18 years Coronary heart disease among adults aged ≥18 years Diagnosed diabetes among adults aged >=18 years Mental health not good for >=14 days among adults aged >=18 years Physical health not good for >=14 days among adults aged >=18 years All teeth lost among adults aged >=65 years Stroke among adults aged >=18 years
Prevention	Visits to doctor for routine checkup within the past year among adults aged ≥18 years Visits to dentist or dental clinic among adults aged ≥18 years Taking medicine for high blood pressure control among adults aged ≥18 years with high blood pressure Cholesterol screening among adults aged ≥18 years Mammography use among women aged 50-74 years Papanicolaou smear use among adult women aged 21-65 years Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50–75 years Older adults aged ≥65 years who are up to date on a core set of clinical preventive services by age and sex
Unhealthy behaviors	Current smoking among adults aged >=18 years No leisure-time physical activity among adults aged >=18 years Obesity among adults aged >=18 years Sleeping less than 7 hours among adults aged >=18 years

6.3.3. Analysis pipeline and Transfer Learning algorithm

The pipeline implemented in our work is described in Figure 6.16. The NAIP NYC image has been subdivided into image square blocks having size of 256x256 pixels, corresponding to a 512 meters edge. It was therefore possible to estimate the value of each of the 24 variables collected by "500 Cities" in each block. During this process, blocks out of the tracts or over the sea have been excluded, thus reducing the dataset. The images have been then processed by a pre-trained deep model, thus extracting the final features for each image. Images are clustered by resorting to k-means clustering, and the clusters (also interpreted with visual inspection) have been associated to the healthcare indexes by statistical analysis.

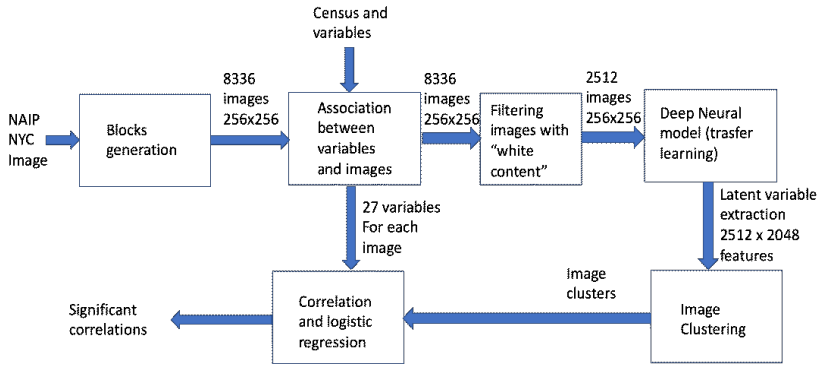


Figure 6.16: The data analysis pipeline created in the study.

After the subdivision of the NAIP image, we obtained 8836 images with dimension 256×256×3 (where 3 indicated the RGB levels), named *tiles*, and we saved all of them in a sequential order starting from the top left corner of the NAIP image. The original image is georeferenced with a world file (see section 4.2), and this georeferencing was kept in the tiles.

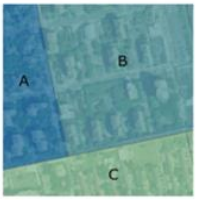
Some of the so-obtained tiles had white areas corresponding to unmapped zones, i.e. zones outside of the city boundary, including the ocean and the rivers. Due to the availability of the vector map of the borders of NYC, we have been able to quantify, for each tile,

the amount of its surface lying inside the borders of the city; we then filtered the original tile set and maintained only those having a minimal overlapping of 90%. After these operations, our final dataset was composed by 2512 images.

The healthcare indexes of the 500 Cities database were then associated to each tile. In order to carry out our analysis, we had to determine the value of the considered variables for each image block, considering that the health variables were taken using the 2166 census tracts as spatial reference, meaning that a given tile overlaps, in general, several tracts. Therefore, we had to implement a simple estimator of the healthcare index of the block, as:

$$hci(block) = \frac{\sum_j w_j hci_j}{\sum_j w_j}$$

where *hci(j)* is the value of the generic health care index for the j^{th} census tract and w_j is the percentage of the image block covered by the mentioned tract. An example is shown in Figures 6.17 and 6.18.



Census tract	SLEEP
A	8,45
B	47,36
C	32,48

$$w_A = \frac{Area\ of\ A}{Area\ of\ image} \times 100$$

$$w_B = \frac{Area\ of\ B}{Area\ of\ image} \times 100$$

$$w_C = \frac{Area\ of\ C}{Area\ of\ image} \times 100$$

$$SLEEP\ (block) = \frac{(w_A \times 8{,}45 + w_B \times 47{,}36 + w_C \times 32{,}48)}{100}$$

Figure 6.17: Example of the quantification of a healthcare index value (SLEEP = percentage of people that declares to be sleeping at least 7 hours per night on average) of a block.

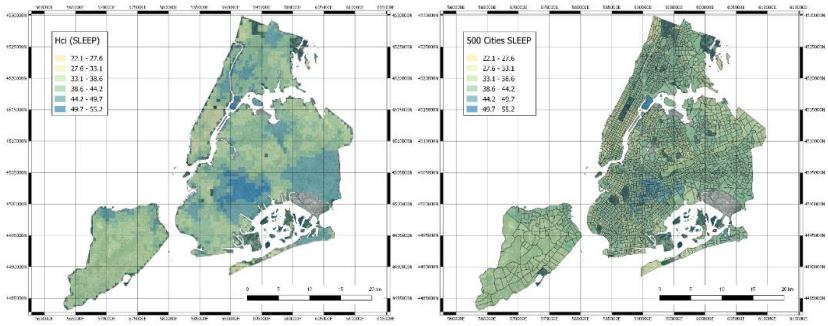


Figure 6.18: Original Census Tracts with the 500 Cities SLEEP index (left hand side) and derived quantification of the healthcare index values for SLEEP variable (right hand side).

Once all these pre-processing operations were terminated, we analyzed the obtained images with a deep learning method, applying the transfer learning concept. As a deep neural network model used for transfer learning, we have selected the network developed for the 2016 Painters by number competition [146]. In such competition the goal was to learn how to discriminate the authors of paintings between 1584 unique painters and starting from a training set of 79433 images; the test set was composed of 23817 images. In this case, a deep neural network model was learned, with 23 layers, mostly convolutional layers with some max pooling layer. The Painters network computes a layer of 2048 latent variables before the final discrimination layer implemented with a soft-max non-linear function. Those latent variables can be used to embed generic images in the latent space. Therefore, using the software Orange (https://orange.biolab.si) and its Python pipeline, we have processed all image blocks with the Painters model, thus obtaining a final data matrix of 2512 examples with 2048 features.

This neural network was selected after testing all those made available by Orange (6 different CNNs with different structures, i.e. Inception v3 [147], VGG-16 [148], VGG-19 [149], Painters, DeepLoc [150], Openface [151]), using the t-SNE algorithm. This algorithm performs a dimensionality reduction projecting multidimensional data into a 2D space, grouping the observation based on their possible similarities in the original space [152]. We tested each neural network and measured their capability of grouping together images with a high percentage of green color,

and the results show that the Painters network was the best performing one. In particular, the features generated by the different deep learning models were mapped onto a bidimensional map using the t-SNE algorithm. Then, the samples with the largest percentage of green color have been selected and the centroid of these samples in the t-SNE space have been computed. Finally, the sum of squared Euclidean distances (SSE) of those samples with the centroid have been computed, and the deep network architecture with the lowest SSE has been thus selected. It should be noted that the information about the percentage of green color of each tile was used only to compare deep learning architectures, it was not used in the following steps of the analysis, as our aim was not to study health in dependence of the number of green areas in the neighborhoods.

6.3.4. Correlation and statistical analysis

The embedding process resulted in a dataset of 2512 images described with 2048 features, correspondent to the laten variables of the CNN. These features have been used to cluster the image blocks by resorting to the well-known K-means clustering algorithm with Euclidean distance. The value of K has been derived with a grid search between 2 and 6 and taking the value that maximize the Silhouette coefficient.

The algorithm found that the images could be divided into 4 clusters that, with visual inspection, seemed to well correspond to different urban landscapes. In detail, Cluster C1 corresponds to green areas, Cluster C2 to residential areas with small houses, Cluster C3 to industrial areas and larger buildings, Cluster C4 to residential with larger buildings. Some examples are shown in Figure 6.19. Cluster analysis clearly shows that the deep neural network model is able to map images in the latent space that share the intuitive notion of similarity that humans may use when they have to classify urban landscape. The method is thus able to automatically cluster similar areas where similar interventions can be planned.

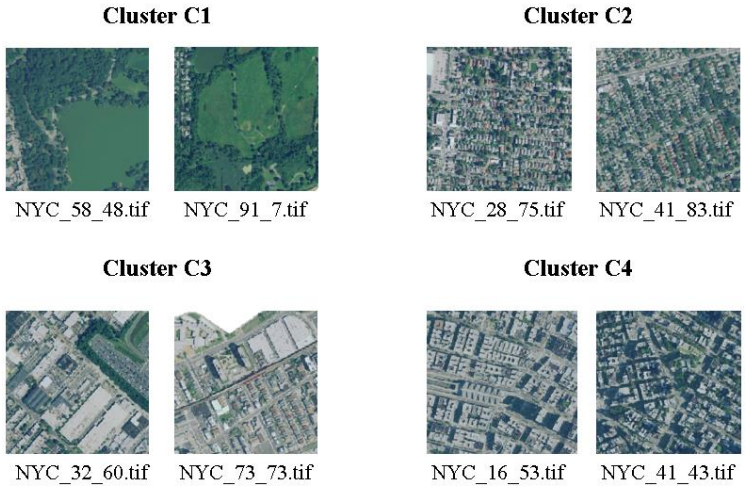
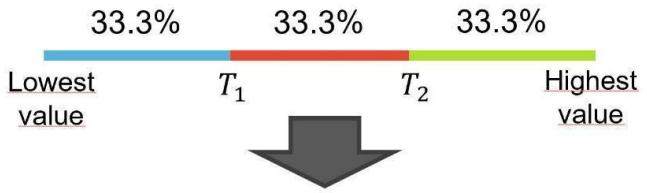


Figure 6.19: A pair of sample images for each cluster that was found.

Once the 4 clusters were found, we performed a few statistical analyses to investigate whether they could be correlated with the health indexes of the 500 Cities database. First, we performed a Chi-squared test over all the 25 variables in order to verify whether they were related to the clusters in a statistically significant way. In order to perform this test, which works on categorical variables, all the continuous variables of the 25 indicators were discretized with the following procedure: each variable was subdivided into 3 categories so that each category contains 1/3 of the total number of observations; after that, a contingency table with dimension 4×3 was created (where 4 is number of clusters and 3 are the categories), indicating for each cluster the number of observations that are within the ranges defined by the thresholds obtained with the discretization of the variables. Figure 6.20 shows a representation of this procedure.



Cluster 1	Obs. < T_1	T_1 < Obs. < T_2	Obs. > T_2
Cluster 2	Obs. < T_1	T_1 < Obs. < T_2	Obs. > T_2
Cluster 3	Obs. < T_1	T_1 < Obs. < T_2	Obs. > T_2
Cluster 4	Obs. < T_1	T_1 < Obs. < T_2	Obs. > T_2

Figure 6.20: Graphical representation of how the variables were discretized in order to create contingency tables to perform the Chi-squared test.

For all the 25 variables, the null hypothesis of independence between the clusters and the distribution of the health indexes could be rejected with a P-value lower than 0.05, thus confirming a statistical association between the clusters and the health indicators.

The contingency tables obtained before the Chi-squared tests can also be visually represented as in Figure 6.21, where the distribution of the observations in each category for each cluster is represented with colors for the variable CHOLSCREEN, i.e. the percentage of adults older than 18 that have been screened for high cholesterol levels. In this image, it appears that the propensity of adults to get screened is higher in the green and residential areas than in the industrialized or highly urbanized ones.

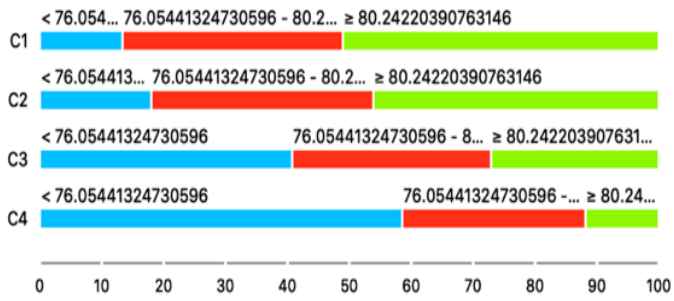


Figure 6.21: The different distributions of Cholesterol screening among adults aged ≥18 years in the different clusters. Inhabitants of cluster C1 have much higher propensity towards screening than those who live in Cluster C4.

Further tests were performed using a set of multinomial logistic regression models treating the four clusters as classes. Two different strategies were used: first, a univariate logistic regression model was fitted for each variable, then a multivariate approach was used to identify the most informative subset of variables with respect to the class taking into account their cross-dependencies.

Of course, being the logistic regression a binomial model, a model built on four classes is actually the result of three sub-models that compare the classes in pairs using one level as baseline. In the context of the analyses presented, Cluster 1 (C1) was considered the baseline class value, while Cluster 2 (C2), Cluster 3 (C3) and Cluster 4 (C4) as the references.

Results from univariate multinomial logistic regression on continuous variables are reported in Table 6.5 and show that all variables tested were significantly associated with at least one cluster (the P-values express the probability to observe "by chance" a difference in terms of variables' distribution between the reference classes compared to the baseline value greater than the effect estimated from data. The null hypothesis is that variables' distribution is the same in each reference cluster compared to the baseline.). The 10 variables showing the strongest statistical association with at least 1 cluster were COREW_Crud, CHOLSCREEN, COREM_Crud, CANCER_Cru, DENTAL_Cru, ACCESS2_Cr, MHLTH_Crud, PAPTEST_Cr, LPA_CrudeP and

COLON_SCRE. Of these, COREW_Crud, CHOLSCREEN, COREM_Crud, CANCER_Cru, DENTAL_Cru, PAPTEST_Cr and COLON_SCRE were characterized by significantly lower values in C2, C3 and C4 compared to C1. Individuals with high values of these variables were less likely to belong to C2, C3 and C4, considering C1 as baseline (OR < 1, p-value < 0.01). On the opposite, subjects with high values of ACCESS2_Cr, MHLTH_Crud and LPA_CrudeP were more likely to belong to C2, C3 and C4 compared to C1 (OR > 1, p-value < 0.001).

A multivariate multinomial logistic regression with a backward stepwise features' selection procedure was then applied to identify the most informative set of variables jointly modulating the probability to belong to the clusters. In this case, 20 variables have been selected (Table 6.6). Of those, five variables have been found to be significant ($p < 0.01$) in all sub-regressions performed by the multinomial model: Colon screening (Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50–75 years), Chronic obstructive pulmonary disease among adults aged ≥ 18 years, High cholesterol among adults aged ≥ 18 years who have been screened in the past 5 years, Chronic kidney disease among adults aged ≥ 18 years and finally Stroke among adults aged ≥ 18 years. Compared to subjects in C1, individuals within C2, C3 and C4 were characterized by significantly higher values of colon screening, higher cholesterol and stroke levels (OR > 2, p-value < 0.001) while lower values of Chronic obstructive pulmonary disease and chronic kidney disease (OR < 0.25, p-value < 0.01).

In general, cluster C1, which is the one that groups green areas, has consistently better prevention and health indicators, but worse sleeping indexes and leisure time. Overall, there is a gradient with all indexes moving from cluster C1, to C2, to C3 and finally to C4, which are the residential areas with large buildings.

Table 6.5: Results from univariate multinomial logistic regression using continuous variables. Variable = variable included in the model; OR = odds ratio expressing the risk to belong to each reference cluster compared to the risk to belong to the baseline cluster C1 by 1 unit increase of each variable; SE = standard error; p = p-value.

	C2 vs. C1		C3 vs. C1		C4 vs. C1	
Variable	**OR (SE)**	**p**	**OR (SE)**	**P**	**OR (SE)**	**p**
ACCESS2_Cr	1.07 (0.01)	<0.001	1.15 (0.01)	<0.001	1.17 (0.01)	<0.001
BINGE_Crud	0.81 (0.02)	<0.001	0.91 (0.02)	<0.001	1.01 (0.02)	0.577
BPHIGH_Cru	1.06 (0.01)	<0.001	1.04 (0.01)	0.004	0.98 (0.01)	0.092
BPMED_Crud	1.09 (0.02)	<0.001	0.91 (0.02)	<0.001	0.87 (0.02)	<0.001
CANCER_Cru	0.86 (0.04)	<0.001	0.6 (0.05)	<0.001	0.47 (0.05)	<0.001
CASTHMA_Cr	1.35 (0.05)	<0.001	1.61 (0.05)	<0.001	1.77 (0.05)	<0.001
CHD_CrudeP	1.04 (0.04)	0.374	0.98 (0.05)	0.660	0.88 (0.05)	0.005
CHECKUP_Cr	1.08 (0.02)	<0.001	0.94 (0.02)	0.001	0.8 (0.02)	<0.001
CHOLSCREEN	0.95 (0.02)	0.001	0.82 (0.02)	<0.001	0.74 (0.02)	<0.001
COLON_SCRE	0.97 (0.01)	0.002	0.9 (0.01)	<0.001	0.88 (0.01)	<0.001
COPD_Crude	1.17 (0.04)	<0.001	1.25 (0.05)	<0.001	1.18 (0.04)	<0.001
COREM_Crud	0.88 (0.02)	<0.001	0.81 (0.02)	<0.001	0.77 (0.02)	<0.001
COREW_Crud	0.88 (0.01)	<0.001	0.82 (0.01)	<0.001	0.77 (0.01)	<0.001
CSMOKING_C	1.06 (0.02)	0.001	1.18 (0.02)	<0.001	1.18 (0.02)	<0.001
DENTAL_Cru	0.94 (0.01)	<0.001	0.9 (0.01)	<0.001	0.9 (0.01)	<0.001
DIABETES_C	1.18 (0.02)	<0.001	1.25 (0.03)	<0.001	1.24 (0.02)	<0.001
HIGHCHOL_C	1.02 (0.02)	0.306	0.92 (0.02)	<0.001	0.86 (0.02)	<0.001
KIDNEY_Cru	2.18 (0.14)	<0.001	2.82 (0.16)	<0.001	2.96 (0.15)	<0.001
X.LPA_CrudeP	1.09 (0.01)	<0.001	1.14 (0.01)	<0.001	1.14 (0.01)	<0.001
MAMMOUSE_C	0.98 (0.02)	0.185	0.9 (0.02)	<0.001	0.83 (0.02)	<0.001
MHLTH_Crud	1.19 (0.03)	<0.001	1.41 (0.03)	<0.001	1.47 (0.03)	<0.001
OBESITY_Cr	1 (0.01)	0.636	1.06 (0.01)	<0.001	1.02 (0.01)	0.084
PAPTEST_Cr	0.85 (0.02)	<0.001	0.84 (0.02)	<0.001	0.81 (0.02)	<0.001
PHLTH_Crud	1.15 (0.02)	<0.001	1.28 (0.02)	<0.001	1.31 (0.02)	<0.001
SLEEP_Crud	1.15 (0.01)	<0.001	1.2 (0.02)	<0.001	1.19 (0.02)	<0.001
STROKE_Cru	1.51 (0.07)	<0.001	1.64 (0.08)	<0.001	1.61 (0.08)	<0.001
TEETHLOST	1.09 (0.01)	<0.001	1.15 (0.01)	<0.001	1.18 (0.01)	<0.001
Green	1 (0)	<0.001	1 (0)	<0.001	1 (0)	<0.001

Table 6.6: Results from multivariate multinomial logistic regression using continuous variables. Variable = variable included in the model; OR = odds ratio; SE = standard error; p = p-value from multivariate multinomial logistic regression.

	C2 vs. C1		C3 vs. C1		C4 vs. C1	
Variable	**OR (SE)**	**p**	**OR (SE)**	**p**	**OR (SE)**	**P**
BPMED_Crud	1.08 (0.12)	0.524	0.65 (0.11)	<0.001	0.81 (0.11)	0.046
CANCER_Cru	0.29 (0.4)	0.002	0.83 (0.41)	0.658	0.84 (0.4)	0.664
CASTHMA_Cr	0.3 (0.48)	0.012	0.58 (0.46)	0.232	7.3 (0.46)	<0.001
CHD_CrudeP	15.53 (0.61)	<0.001	1.18 (0.63)	0.788	1.21 (0.63)	0.764
CHECKUP_Cr	0.73 (0.19)	0.095	0.72 (0.19)	0.083	0.38 (0.17)	<0.001
COLON_SCRE	1.89 (0.1)	<0.001	1.89 (0.1)	<0.001	2.1 (0.12)	<0.001
COPD_Crude	0.12 (0.53)	<0.001	0.21 (0.5)	0.002	0.02 (0.53)	<0.001
COREM_Crud	0.93 (0.12)	0.582	1.34 (0.13)	0.026	1.08 (0.14)	0.574
COREW_Crud	1.1 (0.11)	0.414	0.74 (0.12)	0.012	0.54 (0.13)	<0.001
CSMOKING_C	0.42 (0.24)	<0.001	0.91 (0.22)	0.662	0.7 (0.23)	0.124
HIGHCHOL_C	2.8 (0.17)	<0.001	2.2 (0.19)	<0.001	3.24 (0.2)	<0.001
KIDNEY_Cru	0 (0.76)	<0.001	0.09 (0.75)	0.001	0 (0.77)	<0.001
X.LPA_CrudeP	1.2 (0.12)	0.116	1.37 (0.11)	0.006	0.99 (0.11)	0.925
MAMMOUSE_C	0.91 (0.11)	0.396	0.65 (0.11)	<0.001	0.57 (0.11)	<0.001
MHLTH_Crud	95.21 (0.52)	<0.001	5.03 (0.48)	0.001	1.94 (0.51)	0.195
OBESITY_Cr	0.87 (0.05)	0.004	1.01 (0.05)	0.854	0.81 (0.05)	<0.001
PHLTH_Crud	0.08 (0.52)	<0.001	0.29 (0.47)	0.008	0.62 (0.48)	0.324
SLEEP_Crud	1.59 (0.16)	0.004	1.43 (0.17)	0.031	1.32 (0.16)	0.079
STROKE_Cru	6.86 (0.59)	0.001	42.51 (0.57)	<0.001	384.21 (0.61)	<0.001
TEETHLOST	1.93 (0.15)	<0.001	1.13 (0.15)	0.428	1.56 (0.15)	0.004

6.3.5. Clusters validation

In order to further test the reliability of the found clusters in terms of human interpretation, we performed an additional test based on the human-machine agreement. This concept is widely used in artificial intelligence as a measure of reliability of an

automated classification, following the idea that, in a certain application, the automated process should be able to perform at least as well as a human being, but in a much smaller time. In our case, we pooled the judgement of six people and asked them to rate a few hundreds of images presented to them one by one. These images were taken from the set of tiles that had already been classified by our algorithm, but the result of the classification was unknown to the human raters. Each rater was asked to associate to each image a number ranging from 1 to 5, indicating:

1. Green areas (parks, gardens, nature);
2. Residential areas with small houses;
3. Industrial areas with factories, storage buildings and construction sites;
4. Highly urbanized areas with large buildings;
5. None.

Of the 2512 images classified by the algorithm, 1158 were classified also by our raters, so our comparison was performed on approximately 46% of the entire dataset. Figure 6.22 shows the confusion matrix resulting from this test. In the main diagonal of the matrix it is possible to visualize how many clusters were identified in the same way by both human raters (rows) and the algorithm (columns), on the right side of this matrix two columns reporting the true positive rates (left column) and the false negative rates (right column) are represented, whereas on the bottom the positive predictive value (upper row) and the false discovery rate (lower row) are reported. Looking at these results, we can see that the human-machine agreement is generally moderately high, except for cluster 3, representing the industrial areas. In this case, human raters identified as industrial zones only 43.9% of the images identified in the same way by the algorithm, but, in spite of this, 91.6% of these images were identified as industrial areas also by the algorithm. This result is not unexpected, since industrial areas can be easily mistaken for highly urbanized residential areas, as we can see that most of the human raters classified those images as belonging to this category. Another possible explanation is that the cluster 3 found by the algorithm could not represent specifically industrial areas, but a sort of buffer zone between residential areas and highly urbanized ones, where health indicators tend to worsen

compared to the first ones, but still do not enter in the cluster 4 range of values.

Finally, we analyzed the results using Cohen's Kappa coefficient [153], a metric that measures agreement between two raters taking into consideration also the possibility of the agreement to be occurring by chance. In our case, we obtained Kappa = 0.58, that can be interpreted as moderate agreement, although very close to the substantial agreement threshold, conventionally considered to be 0.6. In particular, it was estimated a random agreement of 0.29 and a maximum possible Kappa of 0.855. Hypothesis testing over the confidence interval of Kappa confirmed these results, rejecting the null hypothesis of random agreement by the two raters with a P-value lower than 1×10^{-4}.

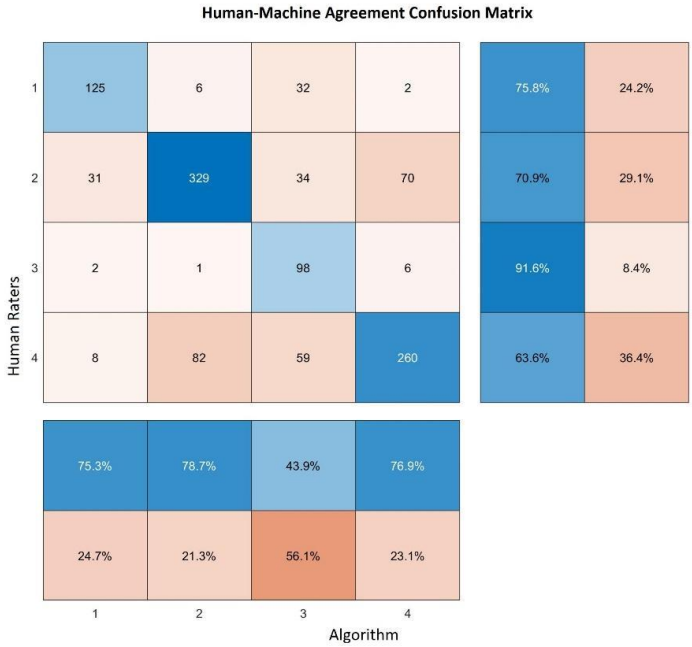


Figure 6.22: confusion matrix of the results of the human-machine comparison.

The found clusters were then represented in the original map of the city, confirming the qualitative evaluation. Figure 6.23 reports

this representation. It is possible to notice that the algorithms clusters together areas that are not necessarily close geographically. It should be noted that there are some areas there were not clustered, this is due to the fact that Some areas in the city are not taken into consideration in the census procedures, since they are mostly uninhabited. Specifically, these areas include the industrial docks, JFK and La Guardia airports, cemeteries and national parks.

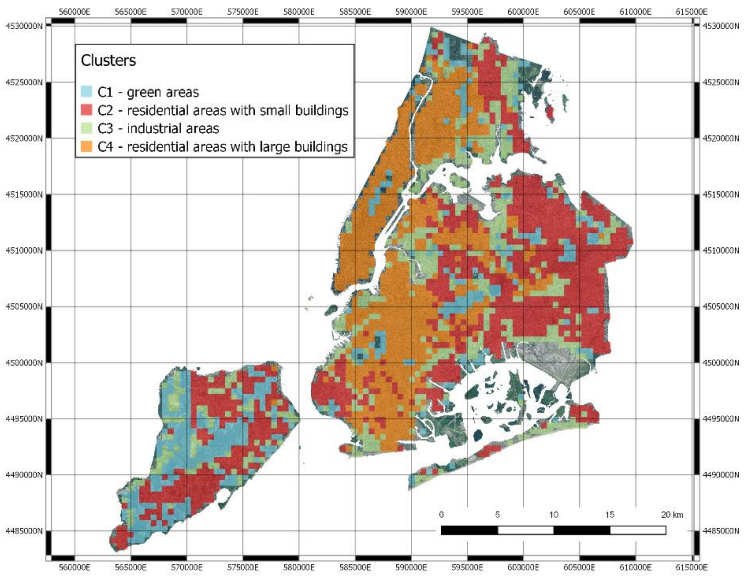


Figure 6.23: the clusters remapped over the original NYC map.

6.3.6. The link between urban landscape and health

The application described in this section is motivated by the assumption that even in a large heterogeneous environment like a big city it could be hypothesized that areas with similar urban structure share also a similar health status, since it is known that the socioeconomic status of a neighborhood is somehow connected to health [59], [154], [155] and mirrored by the urban structure at some point. If that is the case, this information could be used to design intervention strategies even more quickly, since interventions on a specific neighborhood are expected to have

similar effects in other neighborhoods that have the same characteristics.

In this study, we provided proof that correlations between urban structure and health outcomes can be spotted even using something as simple as a satellite image, and we provided an analysis pipeline to find such correlations. We found that in NYC, indubitably one of the most heterogeneous cities in the world, the link between urban landscape and health indicators is particularly strong, and the local presence of green areas such as parks, gardens and nature has an impact on population's health status.

Looking at the detailed results, it is possible to draw the conclusion that areas with the highest percentage of green are also those in which the population tends to have less health complications because prevention is performed more correctly, and people have healthier habits. It should be noted that this apparently simple link may hide a large number of social implications that are the result of several peculiarities of the US social system, where generally speaking health, education and prevention are highly correlated to income and social class, as demonstrated also in the study reported in section 6.2.

Of course, these results do not necessarily show that green areas don't need any kind of interventions, since looking at each single factor some specific criticalities could be spot even in the areas where the general health status is good. For instance, concerning the sleep indicator, which reports the percentage of people that declares to sleep at least 7 hours per night, it is possible to see that in the green areas people tend to be more sleep deprived. This could be the result of longer commute times to go to work or even of a more stressful life caused by mentally demanding jobs. It then seems clear that the interventions should be focused on specific variables for specific areas.

Our work has several implications.

First of all, it shows that deep neural networks designed to encode image data can be successfully reused within transfer learning approaches. Their application to represent urban landscape seems highly effective.

Secondly, in the context of the PULSE project, the capability of finding clusters of similar urban landscape may allow to profile city areas, in which healthcare decision makers may plan similar interventions.

Finally, the combination of urban landscape and healthcare indicators is not only useful to hypothesize the intertwining of these two dimensions, but also to further profile urban areas by finding similar areas with similar behaviors of their inhabitants, thus allowing also lifestyle interventions and more precise and personalized health care policies.

Of course, the analysis has some limitations. First of all, the quantification of the health care indexes in the city blocks have been performed by a weighted averaging of the indexes of the census tracts included in the blocks. The weights are computed taking into account only the spatial overlap and not the actual number of inhabitants of the blocks. Although this issue has already partially dealt with, since census tracts are conventionally designed in order to have similar population densities, and tend to be smaller in highly populated areas, some more precise ways to weigh the results on the population of each areas can be tested. Secondly, the results obtained are probably "proxies" of the wealth of the people living in the different areas. For this reason, results may be representative of specific cities and not generalizable to other ones.

6.4. An extra study: the impact of the Covid-19 Lockdown on air pollution in Pavia, Italy.

Most of the studies related to PULSE that are presented here in this dissertation consider the effect of air pollution on human health, among other things. From the point of view of public health, understanding air pollution gives the possibility to design more effective interventions, and among all the things that need to be understood there are the mechanisms that lead to the production of the most damaging pollutants. The only way to limit the effects of air pollution is limiting the pollution itself, controlling its sources. This is already done in several areas of the planet, for example in the northern Italian region named Po Valley. This area is known to be one of the most polluted areas in Europe [156], with several air pollutants, in particular particulate matter, which often rise to dangerous levels. This is due to an unfortunate combination

of factors such as high population density, high industrial activity and geographic position, as the plain is closed on three sides by the mountains and air tends to stagnate for long periods, especially in the winter months. Traffic limitations are often in place during the most polluted days, but the effectiveness of these interventions has been questioned on several occasions.

While public health organizations, together with nonprofit associations and governments are struggling to apply green solutions to contain the increase in pollution levels, a sudden reduction of air pollutant concentration was seen in many countries between the end of the year 2019 and the year 2020, when most of the productive activities throughout the world had to come to an unexpected stop due to the pandemic caused by the new coronavirus named *Covid-19*. This virus generated most likely in Wuhan, China, and quickly spread causing a high number of intense flu-like syndromes and cases of atypical pneumonia. The first studies conducted on this unknown pathogen demonstrated that it had an abnormally high contagious strength and that the percentage of cases that needed hospitalization, artificial ventilation or eventually led to death were significantly higher than the other known influenza viruses [157], thus the virus had to be contained in order to avoid an overload of the healthcare systems. Despite the initial efforts to contain the disease, the virus spread in several Asian countries outside of China, and at the end of February 2020 the first European case unrelated to the Asian outbreak, thus providing evidence of a local transmission, was found in northern Italy, in the city of Codogno. In the following weeks, Italy had a disastrous increase in cases that forced the government to take drastic actions, and Italy was the first western country to apply severe draconian measures and start a general lockdown, with most of the population confined at home and a shutdown of all nonessential productive activities and services.

This unexpected situation created the chance to study how vehicular traffic and factories impact on air pollution on a normal situation, as for several weeks traffic was limited to essential transportation only and factories not producing essential services were closed, but being march a highly variable period weather-wise, results of many studies were controversial [158]. This is true also because air pollution in the Po Valley is a problem typical of the winter months [45], when cold dense air tends to stagnate in the lower layers of the atmosphere, the phenomenon is less frequent in

the warm season as air is less dense and local breezes are more frequent.

Taken the opportunity, we analyzed PM10 and PM2.5 data coming from our sensors network deployed in Pavia, Italy, with a high spatial and temporal resolution, and compared the measurements of a period that goes from the end of February to the beginning of April of 2020 to the same period of 2019 in order to check if the lockdown had an impact on the urban PM pollution, considering the possible effects of confounders such as meteorological conditions.

6.4.1. Data preparation and exploratory analyses

Even if our sensors measure all kinds of particulate matter (PM1, PM2.5 and PM10), only PM10 and PM2.5 have been considered in our analysis, since they are well agreed indexes of air pollution. We considered hourly pollution data in two periods of time: all the hours from February 24th, 2019 at 00:00 CET to April 2nd, 2019 00:00 CEST and the same period in the year 2020, corresponding to the days after the first Covid-19 cases were found in northern Italy, causing the lockdown initiation. Concerning meteorological data, all measurements of average wind speed, maximum wind gusts and air temperature were collected with the same hourly temporal granularity from the official ARPA portal [159]. Out of the 45 sensors at our disposal in Pavia, only 28 were used for this study since some of them were not active yet in 2019.

All the data measure by our sensors were calibrated using the official ARPA data as reference, using a simple linear regression model:

$$y = a \cdot x + b$$

Where y indicates the data coming from our sensor and x those coming from the ARPA monitoring station. Using several measurements with the same temporal granularity, we calculated separate values of the parameters a and b for PM2.5 and PM10. Our sensor showed a moderately high correlation with the ARPA one, specifically the correlation coefficient was 0.83 for PM2.5 and 0.7976 for PM10. Ones the parameters were estimated, we corrected all the measurements of our sensors with the inverse formula of the equation

above. Specifically, calling PM2.5' and PM10' the crude values measured by our sensors, they were scaled as follows:

- PM2.5 = (PM2.5' + 2.4421) / 1.7964
- PM10 = (PM10' + 7.9712) / 1.1499

Once the preprocessing was terminated, we performed an exploratory analysis of how the pollution trends have changed in the different time periods and how weather conditions have influenced this change. Looking at the absolute quantities, a very irregular trend in pollution measurements can be seen during both years (Figure 6.24), suggesting that external factors can modify these measurements. Some of these external factors that are known to influence pollution are wind and temperature. Figure 6.24 shows the average wind speed measured in the same dates when the PM2.5 concentration was detected in the two years (it should be noted that all measurements performed on February 29th 2020 were excluded from this graph). Looking at the two plots, it is notable also through visual inspection that high peaks of wind speed correspond to lower concentrations of PM2.5. The same happens for PM10 (data not shown).

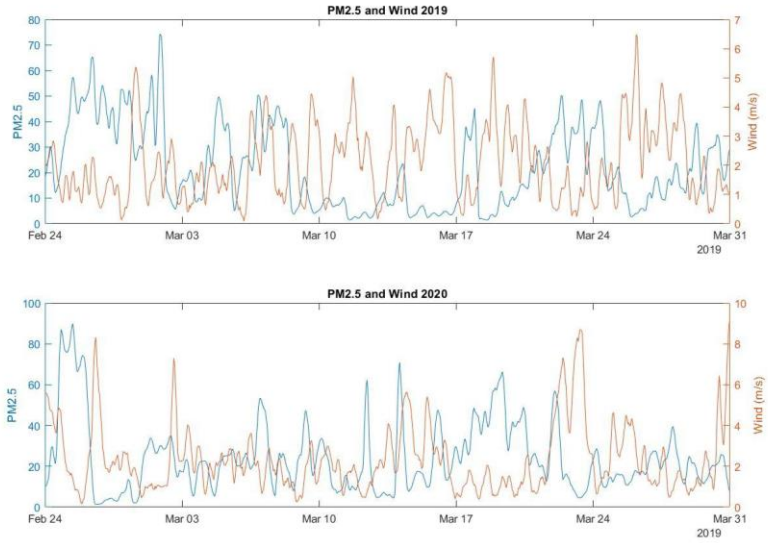


Figure 6.24: PM2.5 concentration and Wind Speed in the two considered periods in 2019 (upper plot) and 2020 (lower plot).

Our preliminary analysis showed that there is a moderate negative correlation between pollution and wind and a weak one between pollution and temperature. Table 6.7 reports the Spearman correlation coefficients and their correspondent 95% confidence intervals.

Table 6.7: Spearman correlation coefficients and 95% confidence interval (95% CI) between particulate matters and wind and temperature.

Pollutant	Correlation with wind (95% CI)	Correlation with temperature (95% CI)
PM10	-0.4222 (-0.4321 : -0.4122)	-0.2408 (-0.2522 : -0.2294)
PM2.5	-0.3627 (-0.3732 : -0.3521)	-0.2091 (-0.2207 : -0.1957)

Performing a quality check on all the sensors' measurements, we noticed that a few sensors presented anomalous readings such as PM2.5 equal to 0 and high values of PM10 detected in the same

timestamp. Although these anomalous measures were taken on isolated moments and thus did not severely affect the distribution of the measurements, those sensors could not be trusted, so data from sensors characterized by at least one (not calibrated) measurement equal to 0 for PM2.5 or PM10 and absolute difference between PM2.5 and PM10 ≥ 9.32 (95^{th} percentile of the absolute difference distribution) were excluded from further analyses due to potential technical issues. The total number of measures available after this quality control criterion was 33,244 deriving from 25 sensors.

Since PM2.5 and PM10 distributions were right-skewed, they were gaussianized by log10 transformation.

6.4.2. Analyses

The relationship between potential confounding factors represented by wind speed (m/s), temperature (°C) and pollutants' concentration was assessed by visual inspection of the scatterplots reported in Figure 6.25.

Plots in Figure 6.25a and Figure 6.25b highlight a negative correlation between wind speed and both log10 PM2.5 and log10 PM10 levels, especially for higher wind speed values. High wind speed values could reduce pollutants concentration, representing a potential confounder when comparing PM levels between 2019 and 2020. Temperature showed no evidence of correlation with pollutants concentration as shown in Figure 6.25c and Figure 6.25d.

Multivariate regression trees were fitted including wind speed and temperature as predictors while Log10 PM2.5 and Log10 PM10 as dependent variable. By visual inspection of the cross-validation results of the unpruned trees it was possible to observe that the first split reduced the relative error of about 14% for both pollutants while further splits caused minor reductions. Thus, by imposing a single split to the regression tree algorithm, a wind speed of 2.45 m/s was identified as the most informative threshold to stratify both Log10 PM2.5 and Log10 PM10 levels.

Pollutants concentration measured when the wind speed was ≥ 2.45 m/s were significantly lower compared to those performed when the wind speed was below the threshold ($p < 0.0001$). It was then decided to focus on a subset of 28,213 measures performed

when the wind speed was below this threshold to avoid its confounding effect.

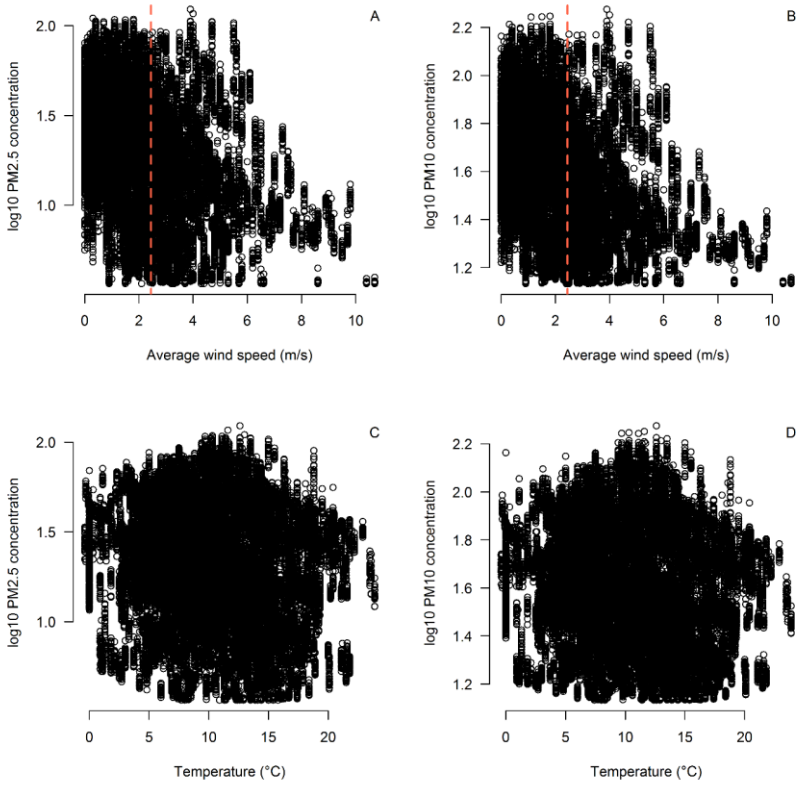


Figure 6.25: Correlation between wind speed, temperature, sensors elevation upon the sea level and pollutants concentration. The vertical dashed line in red indicates the threshold identified by the regression tree.

PM measures were then matched between 2019 and 2020 by sensor, month, day and hour to further reduce the potential impact of confounders when assessing pollutants levels variation between years. A total number of 8,802 paired measures from 20 sensors (year 2019: 4,401, year 2020: 4,401) were included in the analyses based on the matching criteria.

Basic statistical analyses on the mean PM concentrations showed no significant differences between 2019 and 2020. A multivariate linear mixed effects models regression was then fitted to quantify the variation in terms of log10 PM2.5 and log10 PM10 between 2019 and 2020 adjusting by potential confounders represented by wind speed, temperature, weekend/working days during both years and sensors elevation upon the sea level (fixed effect factors). The random effect term was represented by a variable resuming sensor ID and measurement year, month, day and hour.

Results are reported in Table 6.8 and show no statistically significant variation in terms of log10 PM2.5 and log10 PM10 between 2019 and 2020 ($p > 0.05$) accounting for potential confounders.

Table 6.8: Variation in terms of PM2.5 and PM10 between 2019 and 2020 accounting for confounders. Estimate (95% CI) = regression coefficient and 95% CI; p = p-value. Estimates indicate the variation in terms of log10 PM2.5 and log10 PM10.

	Log 10 PM2.5		**Log 10 PM10**	
Variable	**Estimate (95% CI)**	**p**	**Estimate (95% CI)**	**p**
(Intercept)	1.5 (1.43:1.58)	<0.0001	1.72 (1.67:1.78)	<0.0001
Year = 2020	0.01 (0:0.02)	0.0510	0.01 (0:0.02)	0.1168
Working day = yes	0.07 (0.05:0.08)	<0.0001	0.06 (0.05:0.07)	<0.0001
Average wind speed (m/s)	-0.12 (-0.13:-0.11)	<0.0001	-0.09 (-0.1:-0.09)	<0.0001
Temperature (°C)	0.01 (0:0.01)	<0.0001	0.01 (0:0.01)	<0.0001
Sensors' elevation (m)	0 (0:0)	0.0678	0 (0:0)	0.0977

It was then tested for statistically significant variations in terms of Log10 PM2.5 and Log10 PM10 between 2019 and 2020 by daily hours adjusting by confounders. Results are reported in Table 6.9 and evidenced that, compared to the variation in terms of Log10 PM2.5 and Log10 PM10 between 2019 and 2020 observed during from 0:00 am to 6:00 am, pollutants concentration was significantly reduced during the remaining daily hours, except from

8:00 am to 10:00 am when no statistically significant difference was observed.

Interaction analyses were then repeated by sensors characterized by at least 10 measurements, adjusting by confounders. The estimated adjusted variations in the whole sample and by sensor and daily hour are reported in Table 6.9, that shows the results for PM2.5. PM10 results are not reported for brevity, as they showed the same behavior. In particular, results show a general reduction in terms of Log10 PM2.5 and Log10 PM10 between 2019 and 2020 during daily hours (from 10:00 am to 8:00 pm) opposed to an increase during evening, night and early morning (from 8:00 pm to 10:00 am).

The same phenomenon was observed in the median value of PM2.5 and PM10 distribution in the whole sample and by sensor and daily hour.

All the statistical analyses reported were performed by the R statistical software tool. Numeric variables distribution was described by median and interquartile range and by mean and standard deviation. The two tailed t-test for paired samples was applied to test for statistically significant variation between 2019 and 2020. Regression trees were applied by the function *rpart* implemented in the "rpart" package. Linear mixed effects models were applied by the function *lmer* implemented in the R package called "lme4", 95% confidence intervals were estimated by the Wald method. The significance level was set to $\alpha = 0.05$.

Table 6.8: Interaction between year and daily hours in modulating PM2.5 and PM10 variation. Estimate (95% CI) = regression coefficient and 95% CI; p = p-value. Estimates indicate the variation in terms of log10 PM2.5 and log10 PM10.

	Log10 PM2.5		**Log10 PM10**	
Variable	**Estimate (95% CI)**	**p**	**Estimate (95% CI)**	**p**
(Intercept)	1.42 (1.34:1.49)	<0.0000	1.67 (1.61:1.72)	<0.0001
Year = 2020	0.08 (0.06:0.1)	<0.0001	0.05 (0.04:0.07)	<0.0001
hour (6,8]	0.05 (0.02:0.08)	0.0006	0.03 (0:0.05)	0.0173
hour (8,10]	0.03 (-0.01:0.06)	0.1318	0.01 (-0.02:0.03)	0.7018
hour (10,12]	0.02 (-0.02:0.06)	0.2462	0 (-0.03:0.03)	0.9337
hour (12,14]	-0.03 (-0.07:0.01)	0.1826	-0.04 (-0.07:0)	0.0250
hour (14,16]	-0.13 (-0.17:-0.08)	<0.0001	-0.12 (-0.16:-0.09)	<0.0001
hour (16,18]	-0.15 (-0.19:-0.1)	<0.0001	-0.12 (-0.16:-0.09)	<0.0001
hour (18,20]	-0.08 (-0.12:-0.05)	<0.0001	-0.08 (-0.11:-0.05)	<0.0001
hour (20,24]	-0.04 (-0.07:-0.01)	0.0050	-0.05 (-0.07:-0.02)	0.0001
Working day = yes	0.09 (0.08:0.11)	<0.0001	0.08 (0.07:0.09)	<0.0001
Average wind speed (m/s)	-0.13 (-0.14:-0.12)	<0.0001	-0.1 (-0.11:-0.09)	<0.0001
Temperature (°C)	0.02 (0.01:0.02)	<0.0001	0.01 (0.01:0.01)	<0.0001
Sensors' elevation upon the sea level (m)	0 (0:0)	0.0637	0 (0:0)	0.0895
year = 2020:hour (6,8]	-0.07 (-0.1:-0.03)	0.0003	-0.04 (-0.07:-0.02)	0.0011
year = 2020:hour (8,10]	-0.02 (-0.06:0.03)	0.4274	0 (-0.03:0.04)	0.8267
year = 2020:hour (10,12]	-0.17 (-0.22:-0.12)	<0.0001	-0.12 (-0.16:-0.08)	<0.0001
year = 2020:hour (12,14]	-0.23 (-0.29:-0.18)	<0.0001	-0.17 (-0.21:-0.13)	<0.0001
year = 2020:hour (14,16]	-0.12 (-0.17:-0.07)	<0.0001	-0.09 (-0.13:-0.05)	<0.0001
year = 2020:hour (16,18]	-0.08 (-0.13:-0.03)	0.0021	-0.07 (-0.1:-0.03)	0.0013
year = 2020:hour (18,20]	-0.12 (-0.16:-0.08)	<0.0001	-0.09 (-0.12:-0.06)	<0.0001
year = 2020:hour (20,24]	-0.05 (-0.09:-0.01)	0.0109	-0.03 (-0.06:0)	0.0510

Table 6.9: Adjusted mean variation in terms of Log10 PM2.5 between 2019 and 2020 by sensor and daily hours. In green: reductions in terms of pollutants between 2019 and 2020; in red: increase in terms of pollutants between 2019 and 2020.

	Log10 PM2.5 mean variation by daily hour								
Sensor	**(0,6]**	**(6,8]**	**(8,10]**	**(10,12]**	**(12,14]**	**(14,16]**	**(16,18]**	**(18,20]**	**(20,24]**
Overall	**0.077**	**0.012**	**0.059**	**-0.089**	**-0.157**	**-0.042**	**-0.005**	**-0.044**	**0.029**
Asilo Rodari	0.067	0.041	0.089	-0.053	-0.149	0.035	0.016	-0.073	0.048
CREA	0.124	0.050	0.098	-0.059	-0.119	0.030	0.132	0.014	0.044
DICAr	0.088	-0.005	0.070	-0.071	-0.151	-0.063	-0.008	-0.056	0.022
DICAr3	0.118	0.034	0.092	-0.047	-0.134	-0.068	-0.012	-0.031	0.042
EX Palatreves	-0.581	-0.416	-0.675	-0.602	-0.252		-0.019	-0.046	-0.551
Leona house	0.085	0.007	0.035	-0.089	-0.150	-0.039	0.039	-0.035	0.032
P.za Marelli	0.076	0.014	0.070	-0.079	-0.165	-0.047	-0.034	-0.051	0.034
Quartiere Borgo	0.059	0.008	0.063	-0.058	-0.087	-0.007	0.010	-0.009	0.036
Quartiere Nord Est	-0.415	-0.540	-0.866	-0.426	-0.127	0.192	0.148	-0.037	-0.384
Quartiere Ovest	0.099	0.027	0.095	-0.045	-0.139	-0.104	-0.020	-0.011	0.021
Quartiere Scala	0.080	0.003	0.076	-0.070	-0.148	-0.050	-0.009	-0.029	0.028
Sala Broletto	0.072	0.042	0.101	-0.382	-0.224	-0.133	0.041	0.029	-0.031
Scuola Berchet	0.097	0.006	0.027	-0.060	-0.084	-0.010	0.007	-0.069	0.078
Scuola Canna	0.076	-0.003	0.050	-0.098	-0.158	-0.062	-0.035	-0.041	0.042
Via Allende	0.148	0.091	0.122	-0.054	-0.175	-0.043	0.022	-0.030	0.041
Via Corridoni	-0.149	-0.165	-0.218	-0.161	-0.166	-0.109	-0.009	-0.093	-0.002
Via Olevano	0.141	0.065	0.166	-0.174	-0.202	-0.092	-0.009	-0.071	0.030
Via S.Giovannino	0.074	0.029	0.048	-0.075	-0.169	-0.014	-0.013	-0.053	0.054
Via S.Spirito	0.046	-0.016	0.009	-0.124	-0.201	-0.076	-0.066	-0.090	-0.009
Via Tavazzani	0.092	0.022	0.053	-0.083	-0.155	-0.030	0.018	-0.048	0.021

6.4.3. General comments

Reduction of exposure to air pollution is the best prevention strategy that can be applied to reduce the risk of developing respiratory diseases linked to bad air quality. The National Health Ministry of Italy estimated that living in the Po Valley, a large plain region in the north of the country, leads to a significant

reduction of life expectancy [160]. In order to reduce pollution, every year the administrations in the regions and the cities affected take several measures like traffic limitations or regulations on the use of house heating, but the results are often not sufficient, as the sources of pollution are numerous and the combined effects of the produced pollutants is difficult to predict. One of the things that has been clearly observed and that we confirmed in this paper, is that weather has an important impact in this area, as wind, that blows quite rarely and is hardly ever strong in this region, usually reduces pollution, while cold air increases it.

With the new Covid-19 pandemic, that initially had its European epicenter in northern Italy at the end of winter 2020, an important lockdown was imposed throughout the country, creating the opportunity to study and observe the real impact of these sources on the total air pollution situation. Unlike other areas of the world, a severe reduction of all the pollutants during the lockdown in the Po Valley was not observed, indicating that traffic may not have a huge impact on pollution and that meteorological conditions probably play a role even more important than what was usually thought. This situation created to chance to analyze the case study of Pavia, a city in the middle of the Po Valley.

The results of this study show mainly two things: first, the influence of climatic factors on PM levels is extremely high, second, considering this difference, there has not been a significant reduction in PM levels in 2020 during the lockdown, so pollution levels remained high even without traffic and most productive activities, suggesting that in this particular geographical area, most of the relevant pollution sources are the heating systems in private houses and commercial traffic, and even with only essential services active, the peculiar climatic conditions are able to create dangerous levels of pollution. This enlightens the necessity of wide interventions to mitigate the health risks related to air quality in northern Italy. Our results also showed another interesting phenomenon, i.e. a change in the daily pollution patterns during the lockdown, with a higher level of pollutants during the night opposed to a lower concentration in the day hours. This phenomenon could be related to the increased use of house heating deriving from the fact that people stayed more at home in the considered period in 2020, as the night hours are the coldest ones. Nevertheless, pollution is a very complex phenomenon determined by many other confounders somewhat hard to identify.

Chapter 7

Interactive Simulation Tools: applications

Chapter 5 introduces the concept of Agent-based Modeling and simulation tools, explaining how they were chosen to be integrated in the PULSE dashboard. In this chapter, some sample simulation models that have been developed using the PULSE data and following its rationale are described.

In detail, two simple models were developed as examples of public health simulation tools, following the spatially enabled studies performed during the project and reported in chapter 6. One of these models was extended including also traffic dynamics in order to create a vast multilayer model that simulates individual risk taking into account the traffic flows in the different neighborhoods of the city and their effects on urban air pollution from a macroscopic point of view, and adding personal risk factors considered at a microscopic point of view to them.

At the end of the chapter, another interactive tool is briefly described, although it does not properly match the features of a simulation tool, it represents another important interactive tool that uses geospatial information to face health problems. In detail, this tool is an air pollution personal exposure calculator.

7.1. Interactive Simulation Tools

In the final part of the pilot phase of the PULSE project, UNIPV developed a few prototypes of ABMs to be integrated in the

PULSE dashboards. These prototypes simulate real-life urban situations and are meant to be a tool for the public health policy makers to explore the effect of possible changes of some variables in the public health panorama of the urban environment. Through these tools, urban planners can answer to "what-if" questions and have an idea of the possible effects of targeted interventions or possible phenomena that could change the public health equilibrium. It should be noted, as said in chapter 5, that ABMs often simulate scenarios using data gathered in a specific contest, thus using limited knowledge that could lead to results that are not always correspondent to what would really happen if the simulated scenario occurred for real. Nevertheless, agent-based simulation often focuses on looking at the trends rather than quantifying the specific results.

Two models are presented in this section, the first model simulates the impact of several environmental and socioeconomic conditions on the asthma hospitalizations rate in East Harlem, New York City, whereas the second model simulates traffic-caused air pollution in Pavia, Italy. Both models are developed using the NetLogo software, specifically designed to build interactive agent-based models.

7.1.1. Simulation of asthma hospitalizations in East Harlem

The first and main ABM that has been developed and then integrated into the PULSE dashboard simulates the trend of asthma hospitalizations in East Harlem, i.e. a neighborhood located in the upper part of Manhattan, in New York City. According to our results of the spatial enablement study presented in section 6.2, this neighborhood is one of the most affected by asthma hospitalizations, together with the confining south Bronx area, so it was selected as a test site for the implementation of this ABM example, also because the square road network of upper Manhattan is easy to visualize.

This model is based on the asthma hospitalizations study presented in section 6.2, where the Geographically Weighted Regression (GWR) algorithm was used to explore the relation between asthma hospitalizations and a number of environmental,

socioeconomic and demographic factors in the different areas of the city.

More precisely, our ABM (Figure 7.1) is based on real GIS data referred to part of East Harlem, with the following boundaries: Malcolm X Boulevard on the West, Tito Puente Way (E 100th Street) on the South, the FDR Drive on the East and E 126th Street on the North. The background of the model is created using GIS shapefiles that have been directly integrated into NetLogo and contain streets centerlines, sidewalks, buildings and parks. The observer, i.e. the utilizer of the model, can determine the initial population and the traffic density, in order to simulate how a variation of them could influence pollution and exposure to it. The population is expressed in absolute numbers, whereas the traffic density is expressed as the number of vehicles places in a segment between two nodes. The roads are in fact modeled as graphs made of nodes and links, that allow to place the agents representing cars spread throughout the environment and simulate their movement making them move towards the nodes running along the links. The number of nodes that are used to discretize the roads can be chosen using the "node-precision" slider, that divides the roads in a high number of nodes if it is set on a high value and vice versa. It should be noted that a high number of nodes corresponds to a more precise modeling of the real-world roads distribution, but it always leads to longer computational times for the simulation.

The interface features then some sliders where the observer can increase or decrease the percentage of land used for industrial activities, the recycling rate and the obesity rate, in order to simulate the impact of interventions on land use, public services and food policies. All these variables were taken into consideration as covariates for the GWR model presented in section 6.2, together with other demographic and socioeconomic factors that have been excluded from the simulation tools as they are not realistically controllable by a local public health authority. The observer can also set the initial mean and standard deviation of the population's age, a specific age will be given to all people according to a normal distribution. The risk of hospitalizations changes with age (i.e. people under 18 and over 60 are more at risk), plus there is a probability of death that increases dramatically after 75 years of age. Each tick of the model corresponds to 6 months, allowing also to simulate the time needed to see the effects of a possible intervention. Once the observer hits the "Go" button, cars are free

to move on the streets and pollute the area, and people walk in the sidewalks and get exposed to pollution. A plot and some monitors show the current number of hospitalized people, based on the probability computed by the regression model and a 99% discharge rate, derived by the SPARCS data used to build the GWR, as it was seen that most asthma hospitalizations last for a few days and it is very rare for an asthma patient to stay hospitalized for months. The initialized quantities can be changed during the simulation to see the subsequent changes in the hospitalizations trend.

Figure 7.1 shows a capture of the interface of the model. In the central part, the GIS background is displayed with the different layers colored with different colors and where the cars are moving in the streets. The sliders and the monitors are spread all around the graphical representation of the model.

Figure 7.1: Screenshot of the ABM prototype developed to study asthma hospitalizations in East Harlem.

The main underlying model that generates the hospitalizations outcome is a simple linear regression that can be written as follows:

$$Hosp = \beta_0 + \beta_1 \cdot pollution(t) + \beta_2 \cdot land\ use(t) +$$

$$+\beta_3 \cdot obesity(t) + \beta_4 \cdot recycling(t) + \beta_5 \cdot age(t)$$

Where the coefficients are already weighted with the weights found in the GWR study and t represents the current time instant (tick) of the simulation. Air pollution depends on traffic density through a simple linear relation, as it is determined by a linear combination of traffic density and average speed variation of the cars:

$$Pollution = c_1 \cdot traffic\ density + c_2 \cdot speed\ variation$$

Both the relation and the values of the constants are entirely arbitrary, as at the time of the creation of the prototype there were not enough data to quantify the real impact of vehicular traffic on air pollution in NYC. Nonetheless, as already said, the main purpose of this kind of simulation is the visualization of trends rather than the real quantification of the variables.

When the model runs, pollution spreads from the roads following a normal distribution centered in the street centerline that depends on the number of vehicles on the road and the distance from the centerline.

7.1.2. Simulation of the pollution trends in Pavia

A second ABM prototype was developed in PULSE and simulated the effect of traffic-related pollution in Pavia, Italy taking into account also climatic factors. As explained in section 6.4, air pollution is a serious issue in Pavia and in all its geographical area.

In this model, some shapefiles were gathered and loaded into NetLogo, following the idea of the NYC model presented in section 7.1.1. As the NYC model, these shapefiles represent roads, green areas, buildings and pedestrian areas. The agents are vehicles that are placed on the roads and can move through them when the simulation is active (figure 7.2). When the cars are moving, pollution spreads from the roads following the same law it had in the NYC model, and its spreading can be visualized as a red cloud generating from the roads. The observer can define the traffic density and the speed of the wind. Furthermore, a specific patch can be selected, in order to visualize the level of pollution in a precise point of the environment. Each tick corresponds to one hour.

Running the model and changing the variables, the observer can see how pollution increases with traffic density and tends to decrease when the wind speed is high (figure 7.3). The relation between traffic density and pollution is qualitative and arbitrary also in this case due to a lack of sufficient data on the cause-effect relations between traffic and pollution. Pollution tends to decrease exponentially with the increase of wind.

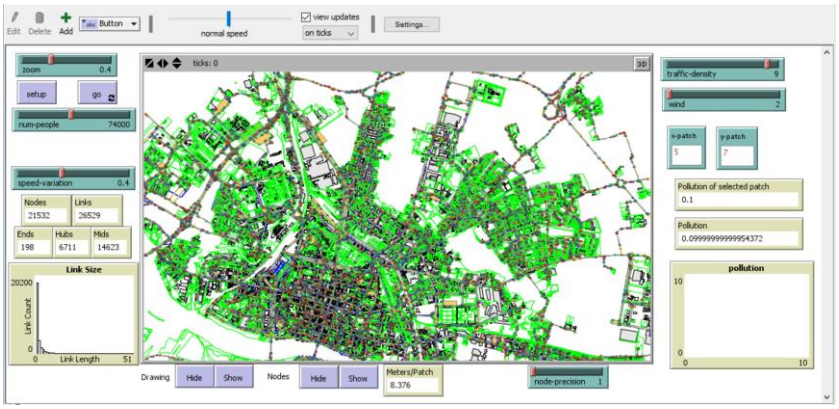


Figure 7.2: Simulation environment of the air quality model prototype of Pavia.

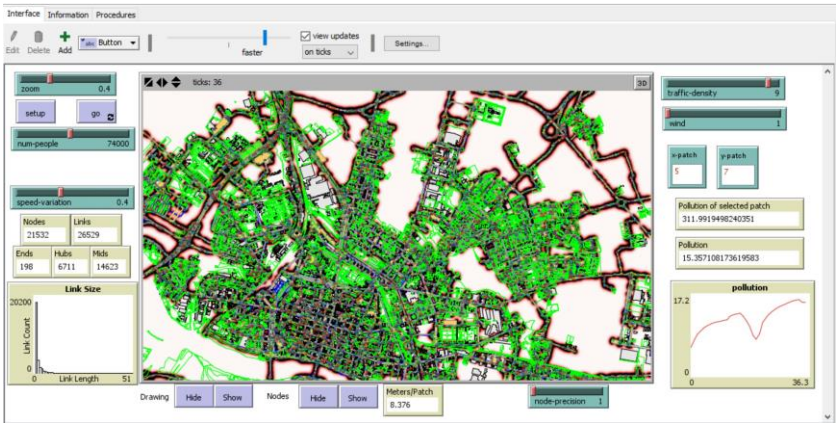


Figure 7.3: Visualization of the model running, with the pollution spreading from the streets and a plot showing its trend during the simulation.

The two prototypes described in this section and the previous one show how ABMs can be a promising tool for public health, as they can be used to simulate several different scenarios and explore the possibility of performing targeted interventions, even if the results can be imprecise from a quantitative point of view.

7.2. A multilayer simulation model for asthma hospitalizations in New York

As explained in chapter 5, traffic modeling can be performed in three ways: creating macroscopic models that consider the effect of aggregated traffic dynamics in one region, creating microscopic models that consider each single vehicle and creating mesoscopic models, that represent a mix of the two. Traffic is an important public health variable as it influences a lot of different health outcomes and it is relatively simple to control with targeted interventions. Not having at our disposal enough data to create an accurate microscopic model, we decided to adopt the macroscopic one presented in section 5.2 to create an innovative simulation tool based on a multilayer approach, that combines traffic dynamics with personal exposure elements to simulate the probability of asthma hospitalizations according to a combination of factors as we did with the GWR, but introducing also more realistic traffic control dynamics.

The model presented here has been developed with reference to the borough of Manhattan, New York City, but the approach can be easily replicated for other boroughs and/or cities. Manhattan has been identified as the proper environment for experimenting this new model, as being an island, it is relatively easy to model the external demand as there is a limited number of access points through bridges and tunnels. Despite being only a part of the city, it has an area of 59.1 km^2 and more than 1.6 million inhabitants, so it can be considered comparable to a city by itself.

7.2.1. Integration of traffic simulation models

The first layer of the model is represented by the macroscopic regional traffic model introduced in section 5.2, that simulates the flow of vehicles through several regions considering the intended flow, the regions that need to be crossed in order to reach the destinations and the external demand, i.e. vehicles entering the regions from outside the simulation space.

The first operation that we performed was dividing Manhattan into zones, this was done arbitrarily trying to reach a trade-off between territorial and traffic homogeneity and health status of the population. Looking at the results of the GWR in section 6.2, it is clear that the asthma hospitalization rate is generally well distributed across the island, with the exception of the upper-east part, correspondent to the neighborhood of East Harlem, that presents peculiar criticalities.

Looking at a map provided by the NYC Department of Transportation that represents the average number of cars passing through the main roads of the city on an average day, visible in Figure 7.4, traffic appears heavy on the FDR (the belt road that runs on the perimeter of the borough), on the access points and generally on the southern part of the city, between the financial district (Downtown) to and the midtown end of Central Park.

Figure 7.4: a map representing the average daily traffic in New York. It can be noticed that the main highways, the bridges/tunnels and the southern half of Manhattan are the most occupied areas.

Looking at these averaged dynamics and considering the population and health characteristics of the borough, we decided to divide Manhattan into 9 areas mostly homogeneous with respect to road density and population. These areas are shown in Figure 7.5, they have been all numbered and are all described with a characteristic Macroscopic Fundamental Diagram (MFD), defined with a triangular function equivalent to the NFDs defined in section 5.2. The numbered areas have been equivalently defined with the names of the neighborhoods they represent, as follows: 1- Downtown, 2- The Village, 3- Chelsea-Gramercy, 4- Midtown, 5- Upper West Side, 6- Upper East Side, 7- Harlem, 8- East Harlem, 9- Washington Heights. It is notable that the area 8 (East Harlem) is much smaller than the others, since its peculiar health situation makes it interesting to observe singularly even if the road density homogeneity criterion is not encountered.

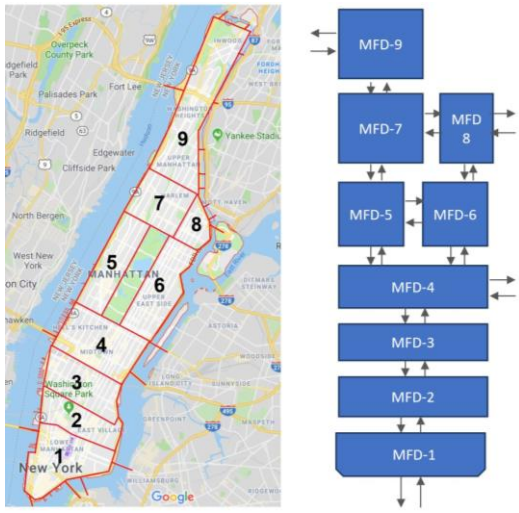


Figure 7.5: on the left side, the map of Manhattan with the subdivisions applied. On the right side, a diagram representing the subdivisions and the areas where the external demand is present.

The peculiar shape of the island makes it relatively easy to model the borders of adjacent zones and to find where the external demand occurs, that is in our case in zones 1, 4, 8 and 9, as they contain bridges or tunnels that connect the island to the other boroughs or to New Jersey.

At this point, the model was implemented in MATLAB and initialized. The main parameter that regulates the directions of the flow inside the network is the *Origin-Destination Matrix* (ODM), i.e. a square matrix with number of rows and columns equal to the number of zones that defines the proportion of vehicles leaving from each zone to head towards the other zones. In detail, the value in the position (o, d) of the matrix indicates the proportion of vehicles exiting the o^{th} zone to go to the d^{th} destination. As a consequence, the sum of each row is equal to one.

Once set the values of the ODM and the initial quantities of vehicles that are located in each zone, the intended flows are defined, but, as shown in the equations of the model in section 5.2, the flows are regulated also by the fact that vehicles that intend to reach a zone non adjacent to the origin one must go through other zones, therefore another parameter to be defined is the paths that vehicles would take to reach their destination. For instance, to go to

zone 2 to zone 4, it is mandatory to pass through zone 3. To define this, the subdivision of the city has been transformed into a graph, visible in Figure 7.6, and we made the assumption that a vehicle that intends to go from one zone to another would take the shortest path possible, i.e. we analyzed the graph and created a set of all the paths that vehicles would take to go from each zone to the others following the shortest road possible.

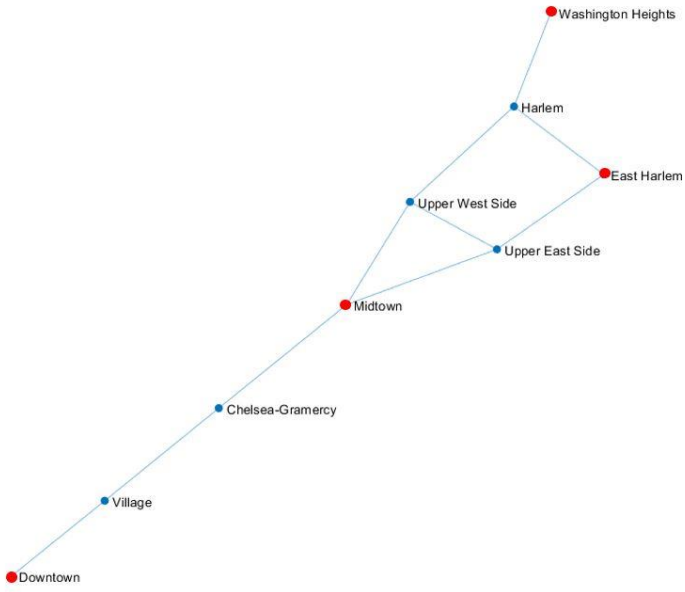


Figure 7.6: graph representation of the zone subdivision of Manhattan. The red dots indicate the areas for which an external demand is defined.

For example, the best path to go from zone 1 to zone 5 is [1,2,3,4,5], whereas the best path to go from zone 1 to zone 6 is [1,2,3,4,6], and the best path to go from zone 7 to zone 4 is [7,5,4]. So for each value of $ODM(o,m)$, one submatrix was defined containing the proportion of the vehicles that left the zone o towards zone d that intend to cross the zones adjacent to o, defining for each time step k the parameter $\rho_{rjd}(k)$ defined in section 5.2, i.e. the density of vehicles moving from the region r to region d passing through the adjacent region j.

After defining the ODM, the simulation time and the initial parameters that regulate the variables to be observed (more details in section 7.2.3), the last parameter that needs to be set is the external demand. In our model, the external demand is defined only for zones 1, 4, 8 and 9, where there are bridges and/or tunnels that connect Manhattan to the rest of land. We defined the "intended" external demand (i.e. the flow that would enter the zone in the absence of capacity constraints) as a linear function defined as follows: let $D_i(0)$ be the number of cars that are supposed to be already entering the zone i at the time stamp $k = 0$ and D_i^{MAX} the absolute quantity of vehicles that can be admitted in the zone during the whole simulation period T, then for each time stamp:

$$D_i(k) = \frac{D_i^{MAX} - (D_{IN}(k) + D_i(0))}{T - (k-1)}$$

Where $D_{IN}(k)$ indicates the quantity of vehicles that has already entered the zone from the beginning of the simulation. In words, the quantity of vehicles entering the zone from outside is constant in each time stamp until the maximum quantity is reached at the end of the simulation.

7.2.2. Integration of health models

Once the traffic simulation dynamics are defined, on top of the vehicles' movements throughout the network, health risk dynamics are added. In particular, in this model the health risk dynamics implemented is the combination of factors that lead to the risk of asthma hospitalizations as found in the GWR model illustrated in section 6.2. The idea is that starting from the macroscopic simulation of traffic, the utilizer of the simulation tool can focus on the population of a single area and study how the asthma risk of an inhabitant exposed to certain factors can change through time with the changes in pollution caused by the traffic variations.

To this end, the first thing to analyze is the relation between traffic and pollution. This is not an easy task since it involves the necessity of a quantitative measure that allows to calculate the quantity of pollutants produced by each vehicle, which is not trivial as the factors that need to be taken into account are very numerous (type and size of the vehicle, combustion type, speed etc.). In our

model, we referred to the EU emission standards [161], and assumed for simplicity that all vehicles had gasoline engines, since diesel cars became very rare in the US in the last decades. Having done these considerations, we associated to each vehicle the production of 0.005 g of PM2.5 per every km made, neglecting its speed. Since in each zone there are thousands of vehicles that transit every time stamp and the measure we are interested in is the PM2.5 concentration in μg/m^3, we estimated this quantity for each area calculating the total length of the roads and the surface of the zone and considering an air column of 100 m. The concentration of pollutant is therefore estimated as follows:

$$C_{PM2.5} = Q_{PM2.5} \cdot L_r / S \cdot 100$$

Where $C_{PM2.5}$ (μg/m^3) is the concentration of PM2.5 in the area, $Q_{PM2.5}$ (μg) is the quantity of pollution produced by all the vehicles, L_r (m) is the total length of the roads in the area, S (m^2) is the total surface of the area and 100 denotes the 100 m air column that was arbitrarily chosen. The measures of L_r and S used for the nine areas are reported in Table 7.1.

Once the quantity of pollution is estimated, the effect on asthma risk is computed using the coefficients obtained with the GWR algorithm, together with the effect of other factors such as age, ethnicity, percentage of land used for industrial activity, obesity, poverty and recycling rate of the neighborhood. According to the algorithm presented in section 6.2, all these factors have an important influence on asthma hospitalizations and the way they influence them changes throughout the city, so it was possible to define different coefficients for the different zones according to the weights found by the algorithm. The asthma risk is then computed as the linear combination of the eight variables listed in Table 7.2 multiplied by the coefficients shown in the same table, that are calculated with respect to the weights found with the algorithm in section 6.2. During the simulation, the pollution value changes due to the traffic flows, leading to a fluctuation of the asthma risk in the course of the simulation as well.

Table 7.1: Road length and surface of each one of the nine areas, used to estimate the pollution level.

Zone	Total Road Length (L_r)	Total Surface (S)
1 - Downtown	70.57 km	6.38 km^2
2 - Village	84.66 km	3.78 km^2
3 – Chelsea-Gramercy	83.98 km	5.032 km^2
4 - Midtown	112.16 km	6.43 km^2
5 – Upper West Side	112.22 km	6.063 km^2
6 – Upper East Side	81.66 km	4.64 km^2
7 - Harlem	94.97 km	5.015 km^2
8 – East Harlem	44.60 km	2.083 km^2
9 – Washington Heights	113.85 km	7.33 km^2

Table 7.2: Coefficients resulting from the GWR algorithm for each considered variable and each zone.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7	Zone 8	Zone 9
Intercept	49.29	35.94	33.05	1.52	-51.87	-74.32	-102.65	-97.33	16.84
PM2.5	-3.45	-2.51	-2.22	-0.67	2.18	3.86	4.68	5.41	-0.99
Age under 18	-1.61	-1.65	-1.55	-1.38	-0.85	-0.74	0.15	0.20	0.64
Hispanic	0.54	0.42	0.41	0.42	0.29	0.30	-0.12	-0.06	-0.42
Black	0.53	0.55	0.55	0.48	0.20	0.16	-0.07	-0.02	-0.10
Poverty	0.24	0.39	0.34	0.63	1.01	1.14	1.19	1.06	0.63
Industrial land use	-0.09	-0.12	-0.11	-0.16	-0.18	-0.27	-0.05	-0.08	0.53
Recycling rate	0.26	0.29	0.22	0.54	0.81	0.83	0.88	0.43	-1.09
Obesity	0.27	0.53	0.56	0.78	1.38	1.47	2.18	1.90	1.03

7.2.3. Simulation

As the traffic and GWR models are both implemented in MATLAB, the simulation runs in the same environment. The main simulation engine is contained in a function that takes as input a parameters' list where the user can define the time stamp and the temporal granularity of the simulation, the number of vehicles in each area when $k = 0$, the values of the socio-economic covariates (e.g. how old is the patient, whether he/she is black or Hispanic, what is the recycling rate of the zone etc.). With the default settings, each time stamp corresponds to one minute and a simulation lasts one hour, but the main function is built in a way that combines eleven simulations in order to simulate the traffic and asthma risk trends in a working day from 8:00 AM to 7:00 PM. For each hour, a different ODM is adopted and the external demand can change as well, in particular there are three main scenarios:

- For the first three simulations, i.e. from 8:00 AM to 11:00 AM, the ODM is defined in a way that leads most cars to the areas from 1 to 4, where most offices and tourist attractions are. The external demand is linear as explained in section 7.2.2 and with a positive sign, i.e. the vehicles that wish to enter Manhattan outnumber the ones that intend to exit the borough.
- During the central hours, from 11:00 AM to 4:00 PM, the ODM is defined randomly, meaning that there is not a clear tendency of the vehicles to move to certain zones rather than others. The external demand is equal to zero, i.e. the number of vehicles that enter Manhattan is balanced by the number of the ones exiting the area.
- The last simulations, corresponding to the day hours from 4:00 PM to 7:00 PM, the ODM is defined in a way that brings most vehicles from the central areas to the zones presenting bridges and tunnels, i.e. most vehicles intend to leave town and go back to the residential areas of the suburbs. The external demand is inverted as the vehicles exiting Manhattan outnumber the number of the ones entering.

Shortly, it could be said that the default parameters allow to perform a simulation of an average working day when traffic tends to be directed towards the heart of the city during the morning and tends to flow away in the evening. The observer can simulate the change of risk of a person with certain characteristics in a specific zone or see how his/her risk would vary depending on the zones. Besides traffic, also the other parameters of the GWR can be changed in order to inspect the consequent variation in the asthma risk. Figure 7.7 shows an example of how the risk varies in all the zones in four different scenarios considering citizens that can be of white or black ethnicity and in a status of poverty.

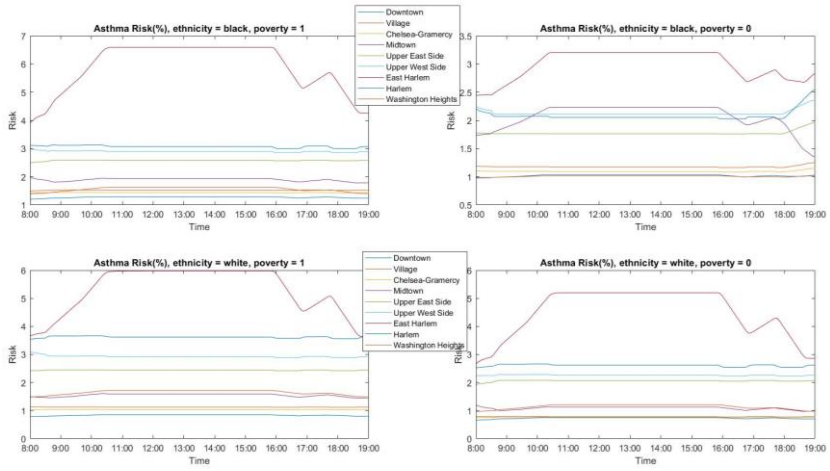


Figure 7.7: Hourly risk in the nine zones in 4 different scenarios: black ethnicity and poverty status (upper left), white ethnicity and poverty status (lower left), black ethnicity without poverty (upper right), white ethnicity without poverty (lower right).

In the example reported in Figure 7.7 it could be noticed that East Harlem has the highest risk in all situations and that the risk is higher during the central hours of the day, when traffic is more intense and leads to an increase in pollution.

The results can indeed be improved with the integration of more detailed real-world traffic data, nevertheless these results show that the model can work as a representation of a real trend of short-term

effects of pollution in the urban area, taking into account also spatial differences.

7.3. Personal Exposure Calculator

Although it cannot be defined a simulation tool per se, another interactive tool developed in the contest of the advanced spatial analytics for collaborative systems explored during PULSE that is worth mentioning is the personal exposure calculator. This tool is based on two main instruments: the dense sensor network deployed in Pavia (described in section 3.3.2) and the GPS tracking functionality of the user's smartphone. The aim of this system is to estimate the exact air pollutants intake of a citizen that moves around in the city in a certain time span, taking into account the local variability of the pollutants' concentration.

Data gathered from the PurpleAir sensors in Pavia showed that both the temporal intra-day variability and the spatial intra-city variability can be significant [162], therefore a high spatial and temporal granularity is necessary to determine the real exposure to pollution of an individual, that could have repercussions on his/her health.

The basis of this system is an interpolated map of the dense sensor network in place in Pavia, that allows to estimate the value of pollution in each point of the city, creating continuous maps from discrete measurements. The methodology used for this operation is the *Gaussian Kernel Interpolation* [163]. This method is based on weighted average: the unknown pollution value z located at the location (x, y, z, t), a point belonging to the 4D space, as time is added to the three spatial coordinates, is calculated as:

$$z = \frac{\sum_j^m \sum_{i=1}^n z_{ij}\, w_i^s w_j^t}{\sum_j^m \sum_{i=1}^n w_i^s w_j^t}$$

where z_{ij} is the pollution level measured by the i-th monitor at the j-th epoch (time stamp when a measurement is taken); usually a certain number of epochs are considered around the selected time: in our case a time window having a semi-width of 2 hours was selected. m is the number of the considered epochs and n is the number of sensors. As the formula highlights, the weight is the

product of the factors w_i^s and w_j^t. Both weight functions are based on a gaussian kernel. The first one is related to space-distance:

$$w_i^s = e^{-\frac{(ds)_i^2}{2\sigma_s^2}}$$

where ds is the spatial distance between the estimation point and the location of the i-th monitor; the function decreases when ds increases; the σ_s parameter controls how quickly the weight decays. The second weight function is related to time-distance and can be written as follows:

$$w_j^t = e^{-\frac{(dt)_j^2}{2\sigma_t^2}}$$

where dt is the time span between the time of the estimation point (x, y, z, t) and the time of the j-th measurement considered. Figures 7.8 and 7.9 show an example of raw measurements of PM10 and the correspondent interpolated continuous map respectively.

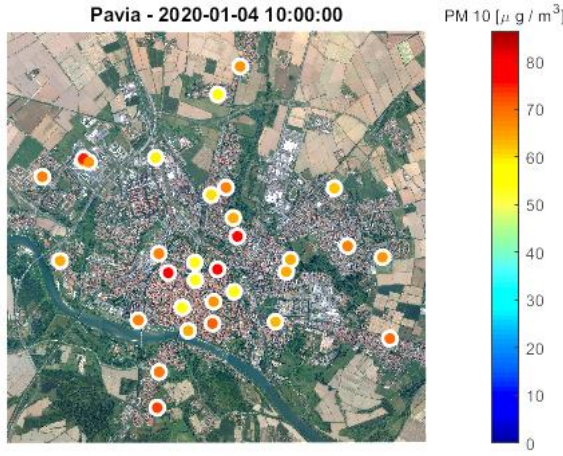


Figure 7.8: measurements of PM10 from all the sensors at a specific timestamp. The colors enlighten a notable local variability.

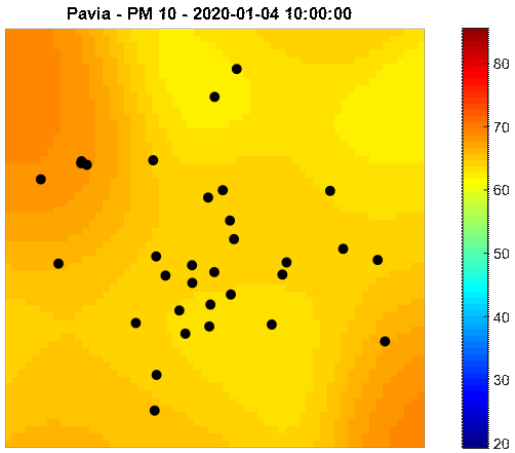


Figure 7.9: interpolated continuous map correspondent to the measurements shown in Figure 7.8.

The users that accept to use this service are tracked by the GPS tracking feature integrated in the Pulsair App, that records their position at regular intervals of 5 seconds.

In order to better estimate the real quantity of pollutants inhaled, physical activity is taken into account as well, as physical exercise leads to an increased intake of air. In order to perform this estimation, speed is determined considering the ration between the user's movement and the distance covered in the unit of time. Following a research in literature [164], different air intake volumes were defined according to the physical activity most likely performed, as shown in Table 7.3

Table 7.3: the breathing model used to estimate the air intake starting from the user's velocity.

Speed [km/h]	Status	# breaths per minute	Air volume per breath [liter]
< 2	At rest	15	0.6
2 - 6	Walking	28	1.8
6 - 15	Running	40	2.5
> 15	Driving	15	0.6

Thanks to this model, the cumulative pollution intake of a user can be precisely estimated. Figures 7.10 and 7.11 show an example of a user tracked during a selected day, in particular the first Figure shows the speed detected during his movements, whereas the second one shows the estimated PM10 intake based on the measured air quality and the physical activity performed. This system has been integrated inside the PULSE system, so that tracks and personal exposure estimations can be visualized both from the App and the WebGIS. The tracks are visible exclusively to the user who produced them and not to other users, public health officials or anyone inside the project consortium.

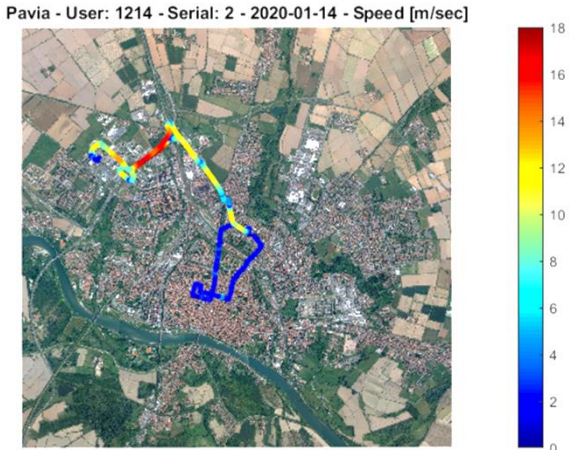


Figure 7.10: cumulative track of a selected user in a specified period, color-coded by speed.

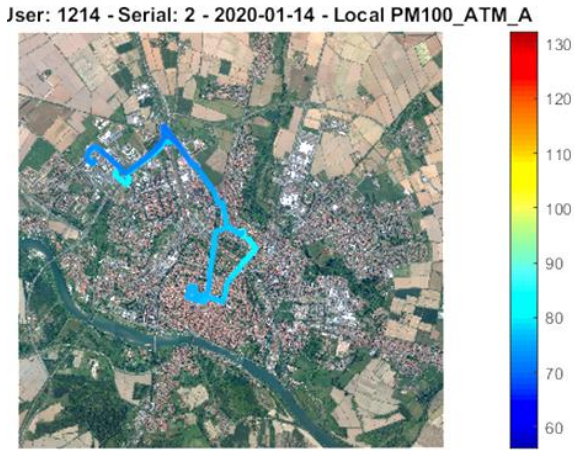


Figure 7.11: Estimated exposure in each point of the track according to the pollution levels computed and the estimated physical activity.

Chapter 8

Conclusions and Future Developments

The recent fast advancements in technology have led to considerable changes in our lifestyle and the consequent worldwide public health panorama. Big cities have become the host of the majority of the world's population and this process appears to have just started, our life habits and the environment are changing causing the rise of new exposure factors we all are exposed to throughout our life, with several repercussions on our health.

These problems are being faced by the scientific community with the establishment of several public health studies and projects that try to assess the entity of the exposures, to better understand the mechanisms that lead to specific health issues and to design possible solutions. Big cities represent the most suitable environment for these kinds of studies and innovations, since they present the highest population and environmental heterogeneity.

The work presented in this dissertation is almost entirely based on one of the latest innovative public health projects that was funded in the last years, named PULSE. PULSE proposes a new approach to face some of the public health issues that are currently rising in the big urban environments, i.e. those mostly connected to air pollution and lifestyle. This approach is based on different levels of intervention: awareness, citizens' behavior and active intervention. The idea is that in order to create a suitable environment for all the citizens, the whole city should cooperate as a system, with awareness of the citizens themselves about the problem and how they could contribute, encouraging them to assume proper behaviors that can aid their health while providing

to the public health policy makers tools to better inspect the situation in the city and organize interventions. To this aim, PULSE created an integrated system that assists all the protagonists of this paradigm, from the App that increases awareness and fosters proper behavior for the citizens, to the dashboards that contain a large set of tools to ease the policy makers' tasks.

Air pollution, for example, is a widely studied topic in medicine and public health, as its negative effect on human health and climate change is well documented. For instance, it has been demonstrated that exposure to most air pollutants has a damaging effect on the airways that lead to a higher risk of respiratory diseases such as asthma. Despite this awareness, there are still several difficulties in finding the best approach to solve these problems, which is mostly diffuse in urban environments, as the problem is often treated without the proper spatial resolution. Most cities have a small number of air monitoring stations and fail to consider some local situations that can generate health issues that are invisible due to the spatial approximation. For this reason, one of the main pillars of PULSE is the increase of spatial granularity in studying and treating these issues. The creation of dense sensor networks such as the one deployed in Pavia allows to find every local criticality and organize more targeted interventions and studies, even during an unexpected situation as the one created by the global Covid-19 pandemic.

The importance of increasing spatial granularity has been one of the fundamentals of the work performed by the author if this thesis and the team from the University of Pavia that worked at the projects that have been presented in this book.

Spatial granularity is not only important for air pollution, but also for all the other exposures that influence our health. This has been demonstrated studying asthma hospitalizations in New York City, where we found that socioeconomic variables can influence the asthma outcomes even more than environmental factors, with their type and level of influence that vary notably across the urban territory, enlightening the necessity to face these issues on a highly local level.

Of course, after studying a problem, also organizing an intervention on it can be a tricky task, especially when the process is long, costly and with unknown outcomes. PULSE provides possible solutions also to this, proposing pipelines to speed up the intervention design procedures clustering urban areas together and

generating simulation tools that allow to explore the possibility of variations in the current urban variables with their effects on health. This has been the second main theme of this book, in which some basic public health simulation tools have been presented and their usefulness for the topic has been widely shown.

Indubitably, the studies presented in this dissertation have some limitations, and the PULSE approach itself presents a number of limitations that require more research and open the road to many future developments. The biggest issue that came out during this project is interoperability: all the projects presented in this thesis are quite city-specific, as they are tuned on the specific social and health environment of the city they are built for. The cause of this issue is partially not resolvable with the current methods, as different cities will always have different populations and environmental contexts, on the other hand part of this problem is due to the lack of standards in the collection and representation of public health data, as different cities and public health authorities collect different data with different standards, and standards sometimes change even within the same city depending on the dataset. This issue has been particularly hard to be dealt with during the creation of the WebGIS, as data from many sources were integrated for seven different cities, but not all the cities were able to provide the same data and formats, spatial and temporal granularity were highly inhomogeneous. Some steps forward in trying to reduce the fragmentation of public health data worldwide should be taken to ease the data integration and analysis processes.

One more important future development for the work presented in this thesis is about the inclusion of human behavior in the health simulation tools developed. It has been demonstrated on several occasions that also mood, happiness, sense of realization etc. can have an important influence on human health, besides wellbeing. PULSE considers all these variables in the risk models used to tune the feedbacks to send to the users, but these topics are not yet treated properly in the projects regarding the dashboard. A proper analysis or simulation tool for a public health policy maker should include also insights on human behavior, to understand also possible reactions to some interventions or changes in wellbeing due to environmental changes in the city. This requires an extra effort, as wellbeing and behavioral models are complex and need the inclusion of many extra studies, nevertheless they could make

an important and useful addition to the tools developed for the PULSE dashboard for the improvement of urban public health.

Besides these limitations, the results of the studies reported in the PhD work presented in this thesis show promising insights on the application of highly spatially enabled methods both to the analysis of health outcomes and to the creation of interactive simulation tools to increase the effectiveness and the velocity of the intervention design process.

References

[1] G. Rosen, *A History of Public Health*. JHU Press, 2015.

[2] E. PERDIGUERO, "Anthropology in public health. Bridging differences in culture and society.," *J. Epidemiol. Community Health*, vol. 55, no. 7, p. 528, Jul. 2001, doi: 10.1136/jech.55.7.528b.

[3] "What is Public Health? | CDC Foundation." http://www.cdcfoundation.org/what-public-health (accessed Sep. 25, 2020).

[4] S. M. Rappaport, "Implications of the exposome for exposure science," *J. Expo. Sci. Environ. Epidemiol.*, vol. 21, no. 1, Art. no. 1, Jan. 2011, doi: 10.1038/jes.2010.50.

[5] P. Vineis, "Exposomics: mathematics meets biology," *Mutagenesis*, vol. 30, no. 6, pp. 719–722, Nov. 2015, doi: 10.1093/mutage/gev068.

[6] C. Anandan, U. Nurmatov, O. C. P. van Schayck, and A. Sheikh, "Is the prevalence of asthma declining? Systematic review of epidemiological studies," *Allergy*, vol. 65, no. 2, pp. 152–167, Feb. 2010, doi: 10.1111/j.1398-9995.2009.02244.x.

[7] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–1053, May 2004, doi: 10.2337/diacare.27.5.1047.

[8] M. J. Pappachan, "Increasing prevalence of lifestyle diseases: high time for action," *Indian J. Med. Res.*, vol. 134, no. 2, pp. 143–145, Aug. 2011.

[9] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Libr.*

Rev., vol. 65, no. 3, pp. 122–135, Jan. 2016, doi: 10.1108/LR-06-2015-0061.

[10] R. Bellazzi, "Big Data and Biomedical Informatics: A Challenging Opportunity," *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 8–13, May 2014, doi: 10.15265/IY-2014-0024.

[11] W. H. Organization and UN-Habitat, *Global report on urban health: equitable healthier cities for sustainable development*. World Health Organization, 2016.

[12] N. Pearce *et al.*, "Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC)," *Thorax*, vol. 62, no. 9, pp. 758–766, Sep. 2007, doi: 10.1136/thx.2006.070169.

[13] X. D. Wang *et al.*, "An increased prevalence of self-reported allergic rhinitis in major Chinese cities from 2005 to 2011," *Allergy*, vol. 71, no. 8, pp. 1170–1180, 2016, doi: 10.1111/all.12874.

[14] "WHO | Asthma," Jun. 29, 2011. https://web.archive.org/web/20110629035454/http://www.who.int/mediacentre/factsheets/fs307/en/ (accessed Sep. 25, 2020).

[15] J. Quirt, K. J. Hildebrand, J. Mazza, F. Noya, and H. Kim, "Asthma," *Allergy Asthma Clin. Immunol. Off. J. Can. Soc. Allergy Clin. Immunol.*, vol. 14, no. Suppl 2, Sep. 2018, doi: 10.1186/s13223-018-0279-0.

[16] F. D. Martinez, "Genes, environments, development and asthma: a reappraisal," *Eur. Respir. J.*, vol. 29, no. 1, pp. 179–184, Jan. 2007, doi: 10.1183/09031936.00087906.

[17] M. Guarnieri and J. R. Balmes, "Outdoor air pollution and asthma," *Lancet*, vol. 383, no. 9928, pp. 1581–1592, May 2014, doi: 10.1016/S0140-6736(14)60617-6.

[18] E. Andersson *et al.*, "Incidence of asthma among workers exposed to sulphur dioxide and other irritant gases," *Eur. Respir. J.*, vol. 27, no. 4, pp. 720–725, Apr. 2006, doi: 10.1183/09031936.06.00034305.

[19] J. Wu, T. Zhong, Y. Zhu, D. Ge, X. Lin, and Q. Li, "Effects of particulate matter (PM) on childhood asthma exacerbation and control in Xiamen, China," *BMC Pediatr.*,

vol. 19, no. 1, p. 194, Jun. 2019, doi: 10.1186/s12887-019-1530-7.

[20] P. Kumar *et al.*, "The rise of low-cost sensing for managing air pollution in cities," *Environ. Int.*, vol. 75, pp. 199–205, Feb. 2015, doi: 10.1016/j.envint.2014.11.019.

[21] F. Karagulian *et al.*, "Review of the Performance of Low-Cost Sensors for Air Quality Monitoring," *Atmosphere*, vol. 10, no. 9, Art. no. 9, Sep. 2019, doi: 10.3390/atmos10090506.

[22] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, "Mapping urban air quality in near real-time using observations from low-cost sensors and model information," *Environ. Int.*, vol. 106, pp. 234–247, Sep. 2017, doi: 10.1016/j.envint.2017.05.005.

[23] M. I. Mead *et al.*, "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmos. Environ.*, vol. 70, pp. 186–203, May 2013, doi: 10.1016/j.atmosenv.2012.11.060.

[24] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Nov. 05, 2020).

[25] S. M. Artham, C. J. Lavie, R. V. Milani, and H. O. Ventura, "The Obesity Paradox: Impact of Obesity on the Prevalence and Prognosis of Cardiovascular Diseases," *Postgrad. Med.*, vol. 120, no. 2, pp. 34–41, Jan. 2008, doi: 10.3810/pgm.2008.07.1788.

[26] H. Ameye and J. Swinnen, "Obesity, income and gender: The changing global relationship," *Glob. Food Secur.*, vol. 23, pp. 267–281, Dec. 2019, doi: 10.1016/j.gfs.2019.09.003.

[27] "What is Diabetes? | NIDDK," *National Institute of Diabetes and Digestive and Kidney Diseases*. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes (accessed Nov. 06, 2020).

[28] L. Kahanovitz, P. M. Sluss, and S. J. Russell, "Type 1 Diabetes – A Clinical Perspective," *Point Care*, vol. 16, no. 1, pp. 37–40, Mar. 2017, doi: 10.1097/POC.0000000000000125.

[29] "Home | ADA." https://www.diabetes.org/ (accessed Nov. 06, 2020).
[30] A. B. Olokoba, O. A. Obateru, and L. B. Olokoba, "Type 2 Diabetes Mellitus: A Review of Current Trends," *Oman Med. J.*, vol. 27, no. 4, pp. 269–273, Jul. 2012, doi: 10.5001/omj.2012.68.
[31] "Prediabetes - Symptoms and causes," *Mayo Clinic*. https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278 (accessed Nov. 06, 2020).
[32] U. Kampmann, L. R. Madsen, G. O. Skajaa, D. S. Iversen, N. Moeller, and P. Ovesen, "Gestational diabetes: A clinical update," *World J. Diabetes*, vol. 6, no. 8, pp. 1065–1072, Jul. 2015, doi: 10.4239/wjd.v6.i8.1065.
[33] "Philadelphia." https://www.citieschangingdiabetes.com/cities/philadelphia.html (accessed Nov. 06, 2020).
[34] E. Nurwanti *et al.*, "Rural–Urban Differences in Dietary Behavior and Obesity: Results of the Riskesdas Study in 10–18-Year-Old Indonesian Children and Adolescents," *Nutrients*, vol. 11, no. 11, Nov. 2019, doi: 10.3390/nu11112813.
[35] C.-C. Hsu *et al.*, "Poverty Increases Type 2 Diabetes Incidence and Inequality of Care Despite Universal Health Coverage," *Diabetes Care*, vol. 35, no. 11, pp. 2286–2292, Nov. 2012, doi: 10.2337/dc11-2052.
[36] P. Paolini, N. D. Blas, S. Copelli, and F. Mercalli, "City4Age: Smart cities for health prevention," in *2016 IEEE International Smart Cities Conference (ISC2)*, Sep. 2016, pp. 1–4, doi: 10.1109/ISC2.2016.7580804.
[37] "Grey and green in Europe: elderly living in urban areas | GRAGE Project | H2020 | CORDIS | European Commission." https://cordis.europa.eu/project/id/645706/it (accessed Nov. 06, 2020).
[38] "Urban GreenUP." https://www.urbangreenup.eu/ (accessed Nov. 06, 2020).
[39] "Integrated Climate forcing and Air pollution Reduction in Urban Systems | ICARUS Project | H2020 | CORDIS | European Commission."

https://cordis.europa.eu/project/id/690105/it (accessed Nov. 06, 2020).
[40] "iSCAPE Project | Improving the Smart Control of Air Pollution in Europe." https://www.iscapeproject.eu/ (accessed Nov. 06, 2020).
[41] "Horizon 2020," *Horizon 2020 - European Commission*. https://ec.europa.eu/programmes/horizon2020/en (accessed Nov. 09, 2020).
[42] E. Teixeira-Lemos, S. Nunes, F. Teixeira, and F. Reis, "Regular physical exercise training assists in preventing type 2 diabetes development: focus on its antioxidant and anti-inflammatory properties," *Cardiovasc. Diabetol.*, vol. 10, no. 1, p. 12, Jan. 2011, doi: 10.1186/1475-2840-10-12.
[43] J. S. Lwebuga-Mukasa and E. Dunn-Georgiou, "The prevalence of asthma in children of elementary school age in Western New York," *J. Urban Health Bull. N. Y. Acad. Med.*, vol. 77, no. 4, pp. 745–761, Dec. 2000, doi: 10.1007/BF02344035.
[44] W. R. W. Lee, "The changing demography of diabetes mellitus in Singapore," *Diabetes Res. Clin. Pract.*, vol. 50, pp. S35–S39, Oct. 2000, doi: 10.1016/S0168-8227(00)00184-4.
[45] A. Bigi, G. Ghermandi, and R. M. Harrison, "Analysis of the air pollution climate at a background site in the Po valley," *J. Environ. Monit.*, vol. 14, no. 2, pp. 552–563, Feb. 2012, doi: 10.1039/C1EM10728C.
[46] K. C. Clarke, "Advances in Geographic Information Systems," *Comput. Environ. Urban Syst.*, vol. 10, no. 3, pp. 175–184, Jan. 1986, doi: 10.1016/0198-9715(86)90006-2.
[47] "Land Surface Temperature," Jul. 31, 2020. https://earthobservatory.nasa.gov/global-maps/MOD_LSTD_M (accessed Nov. 09, 2020).
[48] "Measuring Vegetation (NDVI & EVI)," Aug. 30, 2000. https://earthobservatory.nasa.gov/features/MeasuringVegetation (accessed Nov. 09, 2020).
[49] D. Pala, M. Rocca, and V. Casella, "Advantages and Difficulties of using Spatial Enablement to Support Public Health in Cities: The PULSE Case Study," Nov. 2020, pp. 322–329, Accessed: Nov. 09, 2020. [Online]. Available:

https://www.scitepress.org/Link.aspx?doi=10.5220/0007900003220329.
[50] C. of N. Y. Data NYC Open, "NYC Open Data." http://nycod-wpengine.com/ (accessed Nov. 09, 2020).
[51] "Community Health Survey Trends." https://www1.nyc.gov/site/doh/data/data-sets/community-health-survey.page (accessed Nov. 09, 2020).
[52] "DATA2GOHEALTH.NYC." https://data2gohealth.nyc/ (accessed Oct. 05, 2020).
[53] "500 Cities Project: Local data for better health | Home page | CDC," Dec. 05, 2019. https://www.cdc.gov/500cities/index.htm (accessed Feb. 29, 2020).
[54] "Agenzia Regionale per la Protezione dell'Ambiente della Lombardia." https://www.arpalombardia.it/Pages/ARPA_Home_Page.aspx (accessed Nov. 09, 2020).
[55] "Statewide Planning and Research Cooperative System." https://www.health.ny.gov/statistics/sparcs/ (accessed Oct. 05, 2020).
[56] "DunavNET – Innovative IoT solutions." https://dunavnet.eu/ (accessed Nov. 09, 2020).
[57] "Air Quality Sensors," *PurpleAir*. https://www2.purpleair.com/collections/air-quality-sensors (accessed Nov. 09, 2020).
[58] A. Rajabifard and D. Coleman, "Towards spatial enablement and beyond," GSDI Association Press, 2012.
[59] D. Pala, J. Pagán, E. Parimbelli, M. T. Rocca, R. Bellazzi, and V. Casella, "Spatial Enablement to Support Environmental, Demographic, Socioeconomics, and Health Data Integration and Analysis for Big Cities: A Case Study With Asthma Hospitalizations in New York City," *Front. Med.*, vol. 6, Apr. 2019, doi: 10.3389/fmed.2019.00084.
[60] Q. Liu, M. Deng, Y. Shi, and J. Wang, "A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity," *Comput. Geosci.*, vol. 46, pp. 296–309, Sep. 2012, doi: 10.1016/j.cageo.2011.12.017.

[61] P. M. Atkinson, “Spatially weighted supervised classification for remote sensing,” *Int. J. Appl. Earth Obs. Geoinformation*, vol. 5, no. 4, pp. 277–291, Oct. 2004, doi: 10.1016/j.jag.2004.07.006.

[62] M. E. Hohn, “An Introduction to Applied Geostatistics: by Edward H. Isaaks and R. Mohan Srivastava, 1989, Oxford University Press, New York, 561 p., ISBN 0-19-505012-6, ISBN 0-19-505013-4 (paperback), $55.00 cloth, $35.00 paper (US),” *Comput. Geosci.*, vol. 17, no. 3, pp. 471–473, Jan. 1991, doi: 10.1016/0098-3004(91)90055-I.

[63] A. Hackeloeer, K. Klasing, J. M. Krisp, and L. Meng, “Georeferencing: a review of methods and applications,” *Ann. GIS*, vol. 20, no. 1, pp. 61–69, Jan. 2014, doi: 10.1080/19475683.2013.868826.

[64] R. C. Tryon, *Cluster analysis; correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality,*. Ann Arbor, Mich.: Edwards Brother, Inc., lithoprinters and Publishers, 1939.

[65] C. Wang, Y. Li, S. W. Myint, Q. Zhao, and E. A. Wentz, “Impacts of spatial clustering of urban land cover on land surface temperature across Köppen climate zones in the contiguous United States,” *Landsc. Urban Plan.*, vol. 192, p. 103668, Dec. 2019, doi: 10.1016/j.landurbplan.2019.103668.

[66] K. Ma, L. Guo, and W. Liu, “Investigation of the Spatial Clustering Properties of Seismic Time Series: A Comparative Study from Shallow to Intermediate-Depth Earthquakes,” *Complexity*, Nov. 01, 2018. https://www.hindawi.com/journals/complexity/2018/7169482/ (accessed Nov. 09, 2020).

[67] M. Ergun, H. Uyguçgil, and Ö. Atalik, “Creating a geodemographic classification model within geo-marketing: the case of Eskişehir province,” *Bull. Geogr. Socio-Econ. Ser.*, vol. 47, no. 47, pp. 45–61, Mar. 2020, doi: 10.2478/bog-2020-0003.

[68] T. H. Grubesic, R. Wei, and A. T. Murray, “Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense,” *Ann. Assoc. Am.*

Geogr., vol. 104, no. 6, pp. 1134–1156, Dec. 2014, doi: 10.1080/00045608.2014.958389.

[69] "Accesso di accesso - ArcGIS Online." https://www.arcgis.com/index.html (accessed Nov. 09, 2020).

[70] "Esri: software di mapping GIS, analisi di dati spaziali e location Intelligence." https://www.esri.com/it-it/home (accessed Nov. 09, 2020).

[71] "How Grouping Analysis works—ArcGIS Pro | ArcGIS Desktop." http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-grouping-analysis-works.htm (accessed Feb. 27, 2018).

[72] R. M. AssunÇão, M. C. Neves, G. Câmara, and C. D. C. Freitas, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees," *Int. J. Geogr. Inf. Sci.*, vol. 20, no. 7, pp. 797–811, Aug. 2006, doi: 10.1080/13658810600665111.

[73] F. Santos, V. Graw, and S. Bonilla, "A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon," *PLOS ONE*, vol. 14, no. 12, p. e0226224, Dec. 2019, doi: 10.1371/journal.pone.0226224.

[74] D. P. McMillen, "Geographically Weighted Regression: The Analysis of Spatially Varying Relationships," *Am. J. Agric. Econ.*, vol. 86, no. 2, pp. 554–556, May 2004, doi: 10.1111/j.0002-9092.2004.600_2.x.

[75] Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen, "Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes," *Sensors*, vol. 19, no. 9, May 2019, doi: 10.3390/s19092047.

[76] C.-P. Tsai and T.-L. Lee, "Back-Propagation Neural Network in Tidal-Level Forecasting," *J. Waterw. Port Coast. Ocean Eng.*, vol. 125, no. 4, pp. 195–202, Jul. 1999, doi: 10.1061/(ASCE)0733-950X(1999)125:4(195).

[77] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[78] H. Knutsson and C.- Westin, “Normalized and differential convolution,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 1993, pp. 515–523, doi: 10.1109/CVPR.1993.341081.

[79] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, “Forest Change Detection in Incomplete Satellite Images With Deep Neural Networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017, doi: 10.1109/TGRS.2017.2707528.

[80] M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris, “Oil Spill Identification from Satellite Images Using Deep Neural Networks,” *Remote Sens.*, vol. 11, no. 15, Art. no. 15, Jan. 2019, doi: 10.3390/rs11151762.

[81] R. Cao *et al.*, “Integrating Aerial and Street View Images for Urban Land Use Classification,” *Remote Sens.*, vol. 10, no. 10, Art. no. 10, Oct. 2018, doi: 10.3390/rs10101553.

[82] S. M. and G. V. Stefania Bandini, “Agent Based Modeling and Simulation: An Informatics Perspective,” Oct. 31, 2009. http://jasss.soc.surrey.ac.uk/12/4/4.html (accessed Sep. 28, 2020).

[83] S. Abar, G. K. Theodoropoulos, P. Lemarinier, and G. M. P. O’Hare, “Agent Based Modelling and Simulation tools: A review of the state-of-art software,” *Comput. Sci. Rev.*, vol. 24, pp. 13–33, May 2017, doi: 10.1016/j.cosrev.2017.03.001.

[84] “(PDF) Agent Based Modelling and Simulation tools: A review of the state-of-art software,” *ResearchGate*. https://www.researchgate.net/publication/316002244_Agent_Based_Modelling_and_Simulation_tools_A_review_of_the_state-of-art_software (accessed Sep. 28, 2020).

[85] “NetLogo 6.1.1 User Manual: Programming Guide.” http://ccl.northwestern.edu/netlogo/docs/programming.html#agents (accessed Nov. 09, 2020).

[86] “Introductory Tutorial — Mesa .1 documentation.” https://mesa.readthedocs.io/en/master/tutorials/intro_tutorial.html (accessed Nov. 09, 2020).

[87] B. D. L. Marshall and S. Galea, “Formalizing the Role of Agent-Based Modeling in Causal Inference and

Epidemiology," *Am. J. Epidemiol.*, vol. 181, no. 2, pp. 92–99, Jan. 2015, doi: 10.1093/aje/kwu274.

[88] M. Laskowski, B. C. P. Demianyk, J. Witt, S. N. Mukhi, M. R. Friesen, and R. D. McLeod, "Agent-Based Modeling of the Spread of Influenza-Like Illness in an Emergency Department: A Simulation Study," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 877–889, Nov. 2011, doi: 10.1109/TITB.2011.2163414.

[89] Y. Ge, L. Liu, B. Chen, X. Qiu, and K. Huang, "Agent-Based Modeling for Influenza H1N1 in an Artificial Classroom," *Syst. Eng. Procedia*, vol. 2, pp. 94–104, Jan. 2011, doi: 10.1016/j.sepro.2011.10.012.

[90] J. V. DePasse *et al.*, "Does cost-effectiveness of influenza vaccine choice vary across the U.S.? An agent-based modeling study," *Vaccine*, vol. 35, no. 32, pp. 3974–3981, Jul. 2017, doi: 10.1016/j.vaccine.2017.05.093.

[91] P. Cooley *et al.*, "The Role of Subway Travel in an Influenza Epidemic: A New York City Simulation," *J. Urban Health Bull. N. Y. Acad. Med.*, vol. 88, no. 5, pp. 982–995, Oct. 2011, doi: 10.1007/s11524-011-9603-4.

[92] T. E. Gorochowski *et al.*, "BSim: An Agent-Based Tool for Modeling Bacterial Populations in Systems and Synthetic Biology," *PLOS ONE*, vol. 7, no. 8, p. e42790, Aug. 2012, doi: 10.1371/journal.pone.0042790.

[93] M. Pechoucek and D. Sislak, "Agent-Based Approach to Free-Flight Planning, Control, and Simulation," *IEEE Intell. Syst.*, vol. 24, no. 1, pp. 14–17, Jan. 2009, doi: 10.1109/MIS.2009.1.

[94] S. Fortuna and A. Troisi, "Agent-Based Modeling for the 2D Molecular Self-Organization of Realistic Molecules," *J. Phys. Chem. B*, vol. 114, no. 31, pp. 10151–10159, Aug. 2010, doi: 10.1021/jp103950m.

[95] M. Tracy, M. Cerdá, and K. M. Keyes, "Agent-Based Modeling in Public Health: Current Applications and Future Directions," *Annu. Rev. Public Health*, vol. 39, pp. 77–94, 01 2018, doi: 10.1146/annurev-publhealth-040617-014317.

[96] D. Chao, H. Hashimoto, and N. Kondo, "Dynamic impact of social stratification and social influence on smoking prevalence by gender: An agent-based model,"

Soc. Sci. Med. 1982, vol. 147, pp. 280–287, Dec. 2015, doi: 10.1016/j.socscimed.2015.08.041.

[97] S. T. Cherng, J. Tam, P. Christine, and R. Meza, "Modeling the Effects of E-Cigarettes on Smoking Behavior: Implications for Future Adult Smoking Prevalence," *Epidemiol. Camb. Mass*, vol. 27, no. 6, pp. 819–826, Nov. 2016, doi: 10.1097/EDE.0000000000000497.

[98] Y. Yang, A. V. D. Roux, A. H. Auchincloss, D. A. Rodriguez, and D. G. Brown, "A Spatial Agent-Based Model for the Simulation of Adults' Daily Walking Within a City," *Am. J. Prev. Med.*, vol. 40, no. 3, pp. 353–361, Mar. 2011, doi: 10.1016/j.amepre.2010.11.017.

[99] A. H. Auchincloss, R. L. Riolo, D. G. Brown, J. Cook, and A. V. Diez Roux, "An Agent-Based Model of Income Inequalities in Diet in the Context of Residential Segregation," *Am. J. Prev. Med.*, vol. 40, no. 3, pp. 303–311, Mar. 2011, doi: 10.1016/j.amepre.2010.10.033.

[100] R. Axelrod, "Advancing the Art of Simulation in the Social Sciences," in *Simulating Social Phenomena*, Berlin, Heidelberg, 1997, pp. 21–40, doi: 10.1007/978-3-662-03366-1_2.

[101] D. Westreich, "From Patients to Policy: Population Intervention Effects in Epidemiology," *Epidemiol. Camb. Mass*, vol. 28, no. 4, pp. 525–528, 2017, doi: 10.1097/EDE.0000000000000648.

[102] A. V. Diez Roux, "Invited commentary: The virtual epidemiologist—promise and peril," *Am. J. Epidemiol.*, vol. 181, no. 2, pp. 100–102, Jan. 2015, doi: 10.1093/aje/kwu270.

[103] D. G. Brown, R. Riolo, D. T. Robinson, M. North, and W. Rand, "Spatial process and data models: Toward integration of agent-based models and GIS," *J. Geogr. Syst.*, vol. 7, no. 1, pp. 25–47, May 2005, doi: 10.1007/s10109-005-0148-5.

[104] A. T. Crooks, "Constructing and implementing an agent-based model of residential segregation through vector GIS," *Int. J. Geogr. Inf. Sci.*, vol. 24, no. 5, pp. 661–675, Apr. 2010, doi: 10.1080/13658810903569572.

[105] J. M. Samet, “Traffic, Air Pollution, and Health,” *Inhal. Toxicol.*, vol. 19, no. 12, pp. 1021–1027, Jan. 2007, doi: 10.1080/08958370701533541.

[106] J. O. Klompmaker *et al.*, “Associations of combined exposures to surrounding green, air pollution and traffic noise on mental health,” *Environ. Int.*, vol. 129, pp. 525–537, Aug. 2019, doi: 10.1016/j.envint.2019.05.040.

[107] F. de Souza, O. Verbas, and J. Auld, “Mesoscopic Traffic Flow Model for Agent-Based Simulation,” *Procedia Comput. Sci.*, vol. 151, pp. 858–863, Jan. 2019, doi: 10.1016/j.procs.2019.04.118.

[108] D. Ziemke, S. Metzler, and K. Nagel, “Bicycle traffic and its interaction with motorized traffic in an agent-based transport simulation framework,” *Future Gener. Comput. Syst.*, vol. 97, pp. 30–40, Aug. 2019, doi: 10.1016/j.future.2018.11.005.

[109] Z. H. Khan and T. A. Gulliver, “A macroscopic traffic model for traffic flow harmonization,” *Eur. Transp. Res. Rev.*, vol. 10, no. 2, p. 30, Jun. 2018, doi: 10.1186/s12544-018-0291-y.

[110] A. Jamshidnejad, I. Papamichail, M. Papageorgiou, and B. De Schutter, “A mesoscopic integrated urban traffic flow-emission model,” *Transp. Res. Part C Emerg. Technol.*, vol. 75, pp. 45–83, Feb. 2017, doi: 10.1016/j.trc.2016.11.024.

[111] C. Menelaou, S. Timotheou, P. Kolios, and C. G. Panayiotou, “Joint route guidance and demand management for multi-region traffic networks,” in *2019 18th European Control Conference (ECC)*, Jun. 2019, pp. 2183–2188, doi: 10.23919/ECC.2019.8795819.

[112] C. F. Daganzo, “The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory,” *Transp. Res. Part B Methodol.*, vol. 28, no. 4, pp. 269–287, Aug. 1994, doi: 10.1016/0191-2615(94)90002-7.

[113] “Advantages and Difficulties of using Spatial Enablement to Support Public Health in Cities: The PULSE Case Study | Request PDF,” *ResearchGate*. https://www.researchgate.net/publication/333424838_Advantages_and_Difficulties_of_using_Spatial_Enablement_to_S

upport_Public_Health_in_Cities_The_PULSE_Case_Study (accessed Oct. 01, 2020).
[114] G. P. Anderson, "Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease," *Lancet Lond. Engl.*, vol. 372, no. 9643, pp. 1107–1119, Sep. 2008, doi: 10.1016/S0140-6736(08)61452-X.
[115] B. A. Alhanti, H. H. Chang, A. Winquist, J. A. Mulholland, L. A. Darrow, and S. E. Sarnat, "Ambient air pollution and emergency department visits for asthma: a multi-city assessment of effect modification by age," *J. Expo. Sci. Environ. Epidemiol.*, vol. 26, no. 2, pp. 180–188, Apr. 2016, doi: 10.1038/jes.2015.57.
[116] S. Kant, "Socio-economic dynamics of asthma," *Indian J. Med. Res.*, vol. 138, no. 4, pp. 446–448, Oct. 2013.
[117] K. Shankardass, M. Jerrett, S. D. Dell, R. Foty, and D. Stieb, "Spatial analysis of exposure to traffic-related air pollution at birth and childhood atopic asthma in Toronto, Ontario," *Health Place*, vol. 34, pp. 287–295, Jul. 2015, doi: 10.1016/j.healthplace.2015.06.001.
[118] I. Kloog, B. Ridgway, P. Koutrakis, B. A. Coull, and J. D. Schwartz, "Long- and short-term exposure to PM2.5 and mortality: using novel exposure models," *Epidemiol. Camb. Mass*, vol. 24, no. 4, pp. 555–561, Jul. 2013, doi: 10.1097/EDE.0b013e318294beaa.
[119] C. Butini, "Asthma By The Numbers," *Medium*, Jan. 30, 2018. https://medium.com/asthma-in-the-south-bronx/asthma-by-the-numbers-73553b2c9621 (accessed Oct. 05, 2020).
[120] "Community Health Profiles - NYC Health." https://www1.nyc.gov/site/doh/data/data-publications/profiles.page (accessed Oct. 05, 2020).
[121] F. Silverman, "Asthma and respiratory irritants (ozone)," *Environ. Health Perspect.*, vol. 29, pp. 131–136, Apr. 1979, doi: 10.1289/ehp.7929131.
[122] "New York City Neighborhood Health Atlas," *Tableau Software*. https://public.tableau.com/views/NewYorkCityNeighborhoodHealthAtlas/Home?%3Aembed=y&%3AshowVizHome=n

o&%3Adisplay_count=y&%3Adisplay_static_image=y&%3AbootstrapWhenNotified=true (accessed Oct. 05, 2020).
[123] "Environment & Health Data Portal." http://a816-dohbesp.nyc.gov/IndicatorPublic/Subtopic.aspx?theme_code=2,3&subtopic_id=11 (accessed Oct. 05, 2020).
[124] A. A. Litonjua, V. J. Carey, S. T. Weiss, and D. R. Gold, "Race, socioeconomic factors, and area of residence are associated with asthma prevalence," *Pediatr. Pulmonol.*, vol. 28, no. 6, pp. 394–401, Dec. 1999, doi: 10.1002/(sici)1099-0496(199912)28:6<394::aid-ppul2>3.0.co;2-6.
[125] L. P. Clark, D. B. Millet, and J. D. Marshall, "Changes in Transportation-Related Air Pollution Exposures by Race-Ethnicity and Socioeconomic Status: Outdoor Nitrogen Dioxide in the United States in 2000 and 2010," *Environ. Health Perspect.*, vol. 125, no. 9, p. 097012, 14 2017, doi: 10.1289/EHP959.
[126] J. Maantay, "Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research," *Health Place*, vol. 13, no. 1, pp. 32–56, Mar. 2007, doi: 10.1016/j.healthplace.2005.09.009.
[127] G. L. Larsen, "Differences between adult and childhood asthma," *J. Allergy Clin. Immunol.*, vol. 106, no. 3 Suppl, pp. S153-157, Sep. 2000, doi: 10.1067/mai.2000.109421.
[128] "Asthma - NYC Health." https://www1.nyc.gov/site/doh/health/health-topics/asthma.page (accessed Oct. 06, 2020).
[129] R. Y. Hsia, J. B. Nath, and L. C. Baker, "Emergency department visits by children, adolescents, and young adults in California by insurance status, 2005-2010," *JAMA*, vol. 312, no. 15, pp. 1587–1588, Oct. 2014, doi: 10.1001/jama.2014.9905.
[130] T. Hernandez-Boussard, C. S. Burns, N. E. Wang, L. C. Baker, and B. A. Goldstein, "The Affordable Care Act reduces emergency department use by young adults: evidence from three States," *Health Aff. Proj. Hope*, vol. 33, no. 9, pp. 1648–1654, Sep. 2014, doi: 10.1377/hlthaff.2014.0103.

[131] S. Mohanan, H. Tapp, A. McWilliams, and M. Dulin, "Obesity and asthma: pathophysiology and implications for diagnosis and management in primary care," *Exp. Biol. Med. Maywood NJ*, vol. 239, no. 11, pp. 1531–1540, Nov. 2014, doi: 10.1177/1535370214525302.

[132] J. Fan, S. Li, C. Fan, Z. Bai, and K. Yang, "The impact of PM2.5 on asthma emergency department visits: a systematic review and meta-analysis," *Environ. Sci. Pollut. Res.*, vol. 23, no. 1, pp. 843–850, Jan. 2016, doi: 10.1007/s11356-015-5321-x.

[133] D. Pala *et al.*, "Deep Learning to Unveil Correlations between Urban Landscape and Population Health," *Sensors*, vol. 20, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/s20072105.

[134] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/j.csbj.2014.11.005.

[135] J. Williams, "Semi-Supervised Deep Learning for Monocular Depth Map Prediction," *Max Planck Institute for Intelligent Systems*. https://is.mpg.de (accessed Oct. 08, 2020).

[136] A. C. Krefis, M. Augustin, K. H. Schlünzen, J. Oßenbrügge, and J. Augustin, "How Does the Urban Environment Affect Health and Well-Being? A Systematic Review," *Urban Sci.*, vol. 2, no. 1, Art. no. 1, Mar. 2018, doi: 10.3390/urbansci2010021.

[137] G. T. Blessi, E. Grossi, G. Pieretti, G. Ferilli, and A. Landi, "Cities, the Urban Green Environment, and Individual Subjective Well-Being: The Case of Milan, Italy," *Urban Stud. Res.*, Jan. 2015, Accessed: Oct. 08, 2020. [Online]. Available: https://www.questia.com/library/journal/1G1-462298806/cities-the-urban-green-environment-and-individual.

[138] M. Helbich, Y. Yao, Y. Liu, J. Zhang, P. Liu, and R. Wang, "Using deep learning to examine street view green and blue spaces and their associations with geriatric

depression in Beijing, China," *Environ. Int.*, vol. 126, pp. 107–117, May 2019, doi: 10.1016/j.envint.2019.02.013.

[139] K. Y. Hong, P. O. Pinheiro, L. Minet, M. Hatzopoulou, and S. Weichenthal, "Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks," *Environ. Res.*, vol. 176, p. 108513, 2019, doi: 10.1016/j.envres.2019.05.044.

[140] G. K. Zewdie, D. J. Lary, E. Levetin, and G. F. Garuma, "Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen," *Int. J. Environ. Res. Public. Health*, vol. 16, no. 11, 04 2019, doi: 10.3390/ijerph16111992.

[141] G. Grekousis, "Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis," *Comput. Environ. Urban Syst.*, vol. 74, pp. 244–256, Mar. 2019, doi: 10.1016/j.compenvurbsys.2018.10.008.

[142] S. Sharma, J. E. Ball, B. Tang, D. W. Carruth, M. Doude, and M. A. Islam, "Semantic Segmentation with Transfer Learning for Off-Road Autonomous Driving," *Sensors*, vol. 19, no. 11, Jun. 2019, doi: 10.3390/s19112577.

[143] N. Krieger, "A Century of Census Tracts: Health & the Body Politic (1906–2006)," *J. Urban Health Bull. N. Y. Acad. Med.*, vol. 83, no. 3, pp. 355–361, May 2006, doi: 10.1007/s11524-006-9040-y.

[144] M. F. Domínguez-Berjón, C. Borrell, R. López, and V. Pastor, "Mortality and socioeconomic deprivation in census tracts of an urban setting in Southern Europe," *J. Urban Health Bull. N. Y. Acad. Med.*, vol. 82, no. 2, pp. 225–236, Jun. 2005, doi: 10.1093/jurban/jti047.

[145] "Census tract." https://factfinder.census.gov/help/en/census_tract.htm (accessed Feb. 29, 2020).

[146] K. Team, "Painter by Numbers Competition, 1st Place Winner's Interview: Nejc Ilenič," *Medium*, Dec. 04, 2019. https://medium.com/kaggle-blog/painter-by-numbers-competition-1st-place-winners-interview-nejc-ileni%C4%8D-4eaab5e6ce9d (accessed Feb. 29, 2020).

[147] Xiaoling Xia, Cui Xu, and Bing Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Jun. 2017, pp. 783–787, doi: 10.1109/ICIVC.2017.7984661.

[148] "VGG16 - Convolutional Network for Classification and Detection." https://neurohive.io/en/popular-networks/vgg16/ (accessed Feb. 29, 2020).

[149] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, "Fundus Image Classification Using VGG-19 Architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, Art. no. 1, Jan. 2019, doi: 10.3390/sym11010001.

[150] J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther, "DeepLoc: prediction of protein subcellular localization using deep learning," *Bioinforma. Oxf. Engl.*, vol. 33, no. 21, pp. 3387–3395, Nov. 2017, doi: 10.1093/bioinformatics/btx431.

[151] E. BESENYEI, "OpenFace – Free and open source face recognition with deep neural networks – E&B Software." https://www.eandbsoftware.org/openface-free-and-open-source-face-recognition-with-deep-neural-networks/ (accessed Feb. 29, 2020).

[152] S. Arora, W. Hu, and P. K. Kothari, "An Analysis of the t-SNE Algorithm for Data Visualization," *ArXiv180301768 Cs*, Jun. 2018, Accessed: Feb. 29, 2020. [Online]. Available: http://arxiv.org/abs/1803.01768.

[153] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012.

[154] L. P. Clark, D. B. Millet, and J. D. Marshall, "Changes in Transportation-Related Air Pollution Exposures by Race-Ethnicity and Socioeconomic Status: Outdoor Nitrogen Dioxide in the United States in 2000 and 2010," *Environ. Health Perspect.*, vol. 125, no. 9, p. 097012, 14 2017, doi: 10.1289/EHP959.

[155] A. A. Litonjua, V. J. Carey, S. T. Weiss, and D. R. Gold, "Race, socioeconomic factors, and area of residence are associated with asthma prevalence," *Pediatr. Pulmonol.*, vol. 28, no. 6, pp. 394–401, Dec. 1999, doi:

10.1002/(SICI)1099-0496(199912)28:6<394::AID-PPUL2>3.0.CO;2-6.

[156] L. F. S.r.l, "Nitrogen dioxide and fine particles are threatening Po valley air quality." https://vitesy.com/blog/air-pollution/nitrogen-dioxide-fine-particles-po-valley-air-quality (accessed Jul. 08, 2020).

[157] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, "Features, Evaluation and Treatment Coronavirus (COVID-19)," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.

[158] M. C. Collivignarelli, A. Abbà, G. Bertanza, R. Pedrazzani, P. Ricciardi, and M. Carnevale Miino, "Lockdown for CoViD-2019 in Milan: What are the effects on air quality?," *Sci. Total Environ.*, vol. 732, p. 139280, Aug. 2020, doi: 10.1016/j.scitotenv.2020.139280.

[159] "Richiesta Dati | ARPA Lombardia." https://www.arpalombardia.it/Pages/Aria/Richiesta-Dati.aspx (accessed Oct. 12, 2020).

[160] redazione, "Air pollution shortens the life expectancy of Italians," *Science in the net*, Jun. 04, 2015. http://www.scienceonthenet.eu/content/article/editorial-staff/air-pollution-shortens-life-expectancy-italians/june-2015 (accessed Jul. 08, 2020).

[161] H. Aydogan, M. Hirz, and H. Brunner, "The use and future of biofuels," *Int. J. Soc. Sci.*, vol. 3, no. 4, pp. 12–21, 2014.

[162] V. Casella, M. Franzini, R. Bellazzi, C. Larizza, and D. Pala, "DYNAMIC ASSESSMENT OF PERSONAL EXPOSURE TO AIR POLLUTION FOR EVERYONE: A SMARTPHONE-BASED APPROACH," in *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Aug. 2020, vol. XLIII-B4-2020, pp. 655–663, doi: https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-655-2020.

[163] V. Maz'ya and G. Schmidt, "On approximate approximations using Gaussian kernels," *IMA J. Numer. Anal.*, vol. 16, no. 1, pp. 13–29, Jan. 1996, doi: 10.1093/imanum/16.1.13.

[164] "Your lungs and exercise," *Breathe*, vol. 12, no. 1, pp. 97–100, Mar. 2016, doi: 10.1183/20734735.ELF121.