POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A spatial analysis of railway mobility and its relationship with pandemic spread in Lombardia during 2020

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Greta Galliani**

Student ID: 968854
Advisor: Prof. Francesca Ieva
Co-advisors: Prof. Piercesare Secchi
Academic Year: 2021-22

Abstract

The COVID-19 pandemic in 2020 has impacted the world, affecting health, economy, education, and social behaviour. Much concern was raised about the role of mobility in the disease's spread, with particular attention on public transport. Understanding the relationship between mobility and the pandemic is key to developing effective public health interventions and policy decisions. This thesis investigates the spatial relationship between mobility and the evolution of COVID-19 in 2020, focusing on public railway transport in Lombardia, Italy. The study uses two mobility datasets: one publicly available from Regione Lombardia Open Data Program and the other derived from data by Trenord, a local railway operator. In this framework, a pipeline is developed to estimate dynamical Origin-Destination matrices from Trenord tickets and passenger counts. Global and local Moran indexes are then employed to explore the relationship between a spatial description based on mobility and the mortality rates in 2020. This research starts with the entire Lombardia region and then focuses on the provinces of Brescia, Bergamo and Milano. Results indicate the virus's spread is strongly linked to people's movements. Indeed, the mobility-based spatial weights could identify periods of positive spatial autocorrelation in the epidemic feature and larger clustered areas compared to classical contiguity-based weights. However, no evidence suggests that public railway transport played a more decisive role than overall mobility in epidemic diffusion. Indeed, we never found a positive spatial autocorrelation with the railway mobility-based spatial description in periods of no correlation with overall mobility.

Keywords: mobility, COVID-19, spatial analysis, trip distribution modelling

Abstract in lingua italiana

La pandemia COVID-19 nel 2020 ha avuto un impatto significativo sul mondo, influenzando la salute, l'economia, l'istruzione e la socialità. Sono state sollevate molte preoccupazioni sul ruolo della mobilità nella diffusione della malattia, con particolare attenzione al trasporto pubblico. Comprendere la relazione tra mobilità e pandemia è fondamentale per sviluppare interventi di salute pubblica e restrizioni efficaci. Questa tesi indaga la relazione spaziale tra mobilità ed evoluzione della pandemia COVID-19 nel 2020, concentrandosi sul trasporto pubblico ferroviario in Lombardia, Italia. Lo studio utilizza due set di dati sulla mobilità: uno disponibile pubblicamente dal Programma Open Data di Regione Lombardia e l'altro derivato dai dati di Trenord, l'operatore ferroviario locale. In questo contesto, è stato sviluppato un processo per stimare matrici dinamiche Origine-Destinazione a partire dai biglietti e dai conteggi dei passeggeri forniti da Trenord. L'analisi degli indici di Moran globali e locali esplora la relazione tra una descrizione spaziale basata sulla mobilità e i tassi di mortalità nel 2020. La ricerca parte dall'intera regione Lombardia per poi concentrarsi sulle province di Brescia, Bergamo e Milano. I risultati indicano che la diffusione del virus è fortemente legata agli spostamenti delle persone. Infatti, i pesi spaziali basati sulla mobilità sono in grado di identificare periodi di autocorrelazione spaziale positiva nella risposta epidemica e gruppi spaziali con simili valori dei tassi di mortalità più ampi rispetto ai classici pesi basati sulla contiguità. Tuttavia, nessuna evidenza suggerisce che il trasporto ferroviario pubblico abbia giocato un ruolo più decisivo della mobilità generale nella diffusione dell'epidemia. Infatti, non abbiamo mai riscontrato un'autocorrelazione spaziale positiva con la descrizione spaziale basata sulla mobilità ferroviaria in periodi di assenza di correlazione con la mobilità complessiva.

Parole chiave: mobilità, COVID-19, analisi spaziale, modellizzazione della distribuzione dei viaggi

Contents

Introduction

Motivations

In the early months of 2020, COVID-19 swept the world, causing massive disruptive effects in our societies through loss of lives and measures taken to contain the virus. The first cases were recorded in December 2019 in Wuhan [1], and later COVID-19 spread in Italy, with the first cases detected in Lombardia in February 2020 [2]. The virus quickly circulated throughout the country, and the government was forced to take drastic measures to slow its transmission. School and workplace closures, remote or online teaching, working from home, cancellation of public events and restrictions on public gatherings and meetings, stay-at-home orders, face coverings, restrictions on public transport, and so on are examples of such strategies [3].

These measures resulted in a sharp reduction in people leaving their homes and travelling within cities and across regions. As a result, travel demand fell, and the usage of private vehicles and public transportation dropped dramatically [4]. Local public transit, in particular, was much affected by restrictions, fear of the disease, and lifestyle changes in working and studying habits. Businesses in the field suffered huge losses, reducing their revenues by 50% in 2020 and 42% in 2021. The number of passengers transported has yet to return to pre-pandemic levels in 2023, three years after the COVID-19 outbreak [5].

However, while there are multiple shreds of evidence about the role of mobility in the spread of the disease, the public transport impact still has to be assessed as past research came to different (and sometimes contrasting) conclusions [6, 7].

Assessing the role of mobility on the disease's spread is key to take active interventions in future outbreaks, not only of COVID-19 but of other diseases and not be found unprepared as in the past. In this framework, it is crucial to understand how mobility relates to the epidemic phenomenon and how the restrictions have affected this relationship. The hypothesis that public transport did not have such a prevalent role in epidemic diffusion should be analysed in detail. If confirmed, it calls for a review of some restrictions on its usage for future outbreaks. This could release businesses in the field of huge losses and encourage people not to fear this transportation method, associated with multiple benefits both for the population (less traffic, less expensive, less stressful and safer compared to travelling by private car) and for the planet (much less pollutant than private cars) [8].

This work is motivated as a part of the **CHANCE project**, specifically of the working group devoted to evaluating the effects of health strategies for reducing SARS-CoV-2 risk, aiming to predict the implications of possible future epidemics. Our goal is to assess the role of mobility, particularly public railway mobility, in epidemic diffusion. A collaboration with the railway company **Trenord** has provided some data about railway mobility.

Concerning the data provided by **Trenord**, we will develop a procedure to derive dynamical Origin-Destination (OD) matrices (i.e., matrices describing mobility in a transportation network, whose cells t_{ij} represent the number of trips starting from zone i and ending in zone j in a specific time frame) describing movements by train in the study area by Trenord data. These data will be used to describe in detail the evolution of weekly railway mobility flows in 8 months of 2020.

We will first analyse the relationship between mobility and the epidemic in the Lombardia area, considering mobility data publicly available by Regione Lombardia and modelling the epidemic response with death counts data by ISTAT. Then, we will repeat the spatial analysis in a limited area partially covering the provinces of Brescia, Bergamo and Milano (named BreBeMi for this reason) and assess the role of overall mobility (i.e., mobility by all means of transportation) versus railway mobility, using static mobility data by Regione Lombardia and the Trenord-derived dynamic railway mobility flows.

Our study aims to answer two research questions:

- *Q1: Can we establish a link between the evolution of the pandemic during 2020 and mobility?*
- *Q2: Did railroad mobility play a relevant role in the pandemic diffusion compared to other kinds of mobility?*

Data related to the research

This dissertation studies the epidemic evolution of the pandemic in Italy, modelling it with the death counts from all causes released by ISTAT [9], in analogy with what has been done in other works, such as [10–12]. The disease's spread is analysed together with two kinds of mobility data provided by the Regione Lombardia Open Data Program and by Trenord.

Regione Lombardia Open Data Program

Regione Lombardia initiated its Open Data program in 2012, intending to promote transparency and accountability by publicly making numerous datasets available. The program aims to provide citizens, businesses, and government agencies with valuable information for their queries. The program adheres to guidelines to ensure quality and regular data updates. Moreover, Regione Lombardia offers the Open Data Lombardia portal to local authorities free of charge for publishing their data [13].

Currently, the Open Data portal of Regione Lombardia is recognised as one of the primary reference points in Italy, thanks to the number and quality of data displayed and, most importantly, the advanced functions it offers to various users such as citizens, developers, and researchers.

In addition to collecting datasets from approximately 200 local administrations, the portal includes sections dedicated to specific entities such as the Municipality of Bergamo, Municipality of Monza, Province of Monza and Brianza, Metropolitan City of Milan, and an exclusive section for the Epidemiological Observatory of the Lombardy Region [14].

Trenord

Founded on May 3, 2011, by the two current shareholders, FNM and Trenitalia, Trenord is among Europe's most important local public rail transport companies, in terms of size and service capillarity.

Operating on a 2,000-kilometer network that links 460 stations, Trenord runs over 2,170 trips daily, connecting Lombardia, seven neighbouring provinces (Alessandria, Novara, Parma, Piacenza, Verbano-Cusio-Ossola, Vercelli, Verona), and Malpensa International Airport via the Malpensa Express rail link.

In Lombardy, 77% of municipalities and 92% of citizens have a railway station within a 5-kilometre radius. Trenord here counts more than 550,000 passengers and 2,200 train rides a day [15].

Thesis contributions

The contributions of this dissertation to the present literature regarding the relationship between COVID-19 and mobility can be summarised as follows:

- This dissertation is the first study, to the best of our knowledge, exploring the spatial autocorrelation of an epidemic feature using a spatial description derived from mobility in the form of mobility-based spatial weights. We use global and local Moran indexes to identify periods of positive spatial autocorrelation in the epidemic feature according to the mobility-based spatial description and to identify locations showing clustered behaviour. We detail a procedure to reproduce the spatial analysis and highlight the role of mobility in the epidemic spread. This procedure is entirely general and could be applied to other kinds of areal epidemic features and mobility data in the form of OD matrices in the future.

- This dissertation builds a detailed pipeline to estimate dynamical OD matrices for a limited portion of the Trenord network, applying the Furness method for trip distribution modelling starting from ticket data and passenger counts obtained by the Automatic Passenger Counting system. This study introduces some novelties compared to past works in trip distribution modelling, such as the conversion of tickets and subscriptions into estimated trips, the development of a model to estimate missing counter data, and the application of a procedure to correct zero values in the seed matrices for the Furness algorithm. These innovations' contribution to the estimation pipeline is discussed together with the results of the procedure.

Thesis outline

The remainder of the thesis is organised as follows:

Chapter 1 - Timeline of the COVID-19 pandemic in Italy during 2020 summarises the main events regarding the history of COVID-19 in Italy during 2020.

Chapter 2 - Literature Review analyses past research on the relationship between COVID-19 spread and mobility.

Chapter 3 - Datasets on mobility and epidemic introduces the epidemiological and mobility datasets used in the analysis.

Chapter 4 - Estimation of Trenord Origin-Destination Matrices presents the theoretical notions behind the Furness method for trip distribution modelling. Then, it defines and applies the pipeline to derive weekly OD matrices describing movements by train in a limited portion of the Trenord network consisting of six train lines during eight months of 2020. Finally, the estimated OD matrices are compared with the Regione Lombardia mobility dataset.

Chapter 5 - Spatial Analysis of Mobility and Epidemics details some theoretical notions of spatial data analysis and defines the pipeline applied to analyse the relationship between mobility and epidemic spread in the two areas considered in our work. Then, it shows the results of the spatial analysis of mortality rates in the Lombardia area through 2020, based on the mobility flows described by the Regione Lombardia mobility dataset. Next, it repeats the spatial analysis in the BreBeMi area, adopting a spatial granularity induced by the distribution of stations in the territory. Overall mobility is compared with railway mobility, first repeating the analysis with the Regione Lombardia mobility data in the new area and then plugging the Trenord-derived data about real movements by train in 2020 into the analysis.

1 | Timeline of the COVID-19 pandemic in Italy during 2020

Given the nature of the study, we summarise the main events regarding the history of COVID-19 in Italy during 2020, focusing on the Lombardia region. We describe the outbreak of COVID-19 in the world, the evolution of the pandemic in Italy, and the restrictions adopted by the Italian government, particularly concerning public transport regulations. We will often refer to this brief history throughout the dissertation. The account of the events is taken by [16] unless differently cited. Figure 1.1 shows the key events in the COVID-19 outbreak in Italy in 2020, while Figure 1.2 displays the curves describing new confirmed cases of COVID-19 and the number of deaths from all causes during 2020.

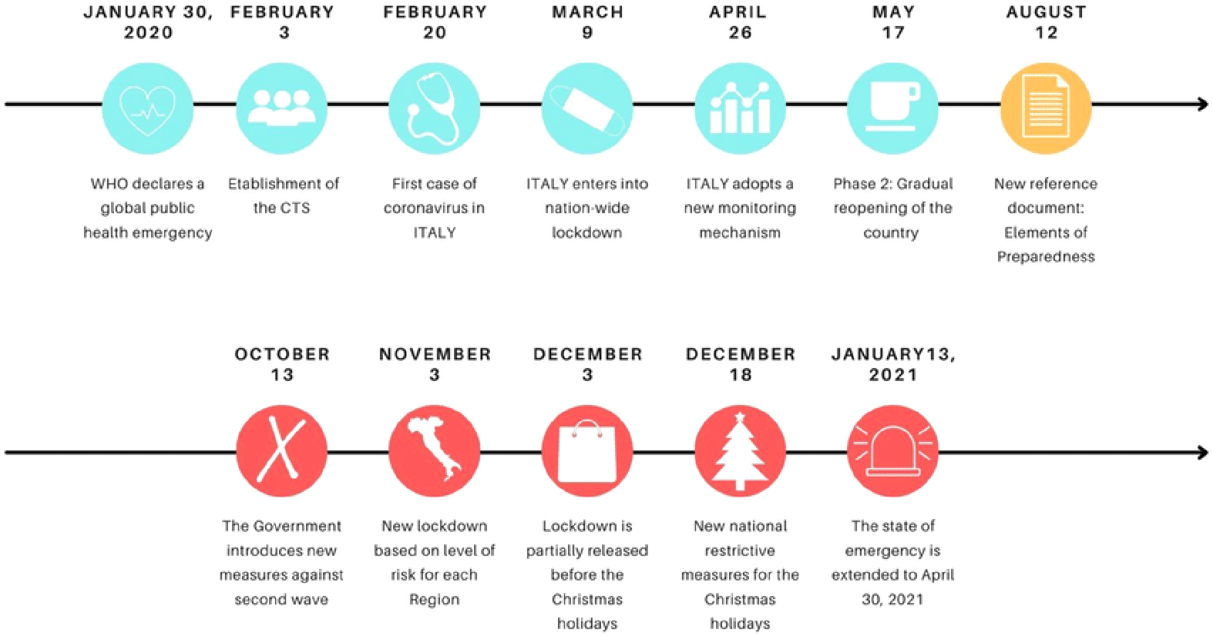
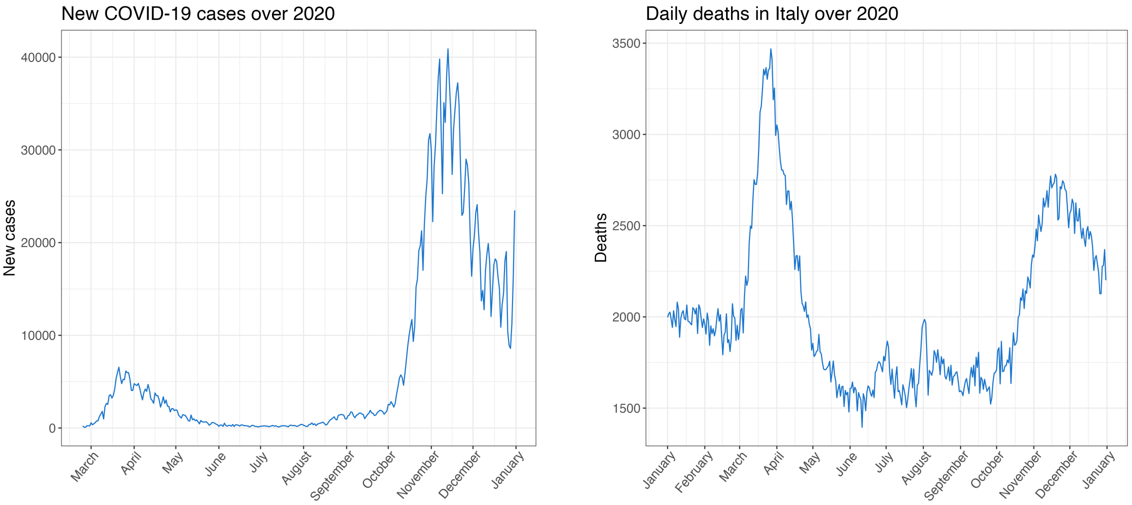

JANUARY 30, 2020
WHO declares a global public health emergency
FEBRUARY 3
Etablishment of the CTS
FEBRUARY 20
First case of coronavirus in ITALY
MARCH 9
ITALY enters into nation-wide lockdown
APRIL 26
ITALY adopts a new monitoring mechanism
MAY 17
Phase 2: Gradual reopening of the country
AUGUST 12
New reference document: Elements of Preparedness
OCTOBER 13
The Government introduces new measures against second wave
NOVEMBER 3
New lockdown based on level of risk for each Region
DECEMBER 3
Lockdown is partially released before the Christmas holidays
DECEMBER 18
New national restrictive measures for the Christmas holidays
JANUARY13, 2021
The state of emergency is extended to April 30, 2021

Figure 1.1: Timeline of the key events and restrictions in the COVID-19 pandemic in Italy [17]

The history of the COVID-19 pandemic started in China: Chinese authorities reported



(a) Daily number of new confirmed cases [18]

(b) Daily number of deaths from all causes [9]

Figure 1.2: Evolution of the curves describing two epidemic indicators referring to COVID-19 in Italy through 2020

to the WHO (World Health Organization) the emergence of several cases of a mysterious pneumonia in December 2019. The epicentre was in Wuhan, a Chinese city of 11 million inhabitants in Hubei province. On January 23, the first lockdown was put into place in Wuhan, followed by other Chinese regions: people could not leave their homes unless strictly necessary and had to wear masks when going out. This was the beginning of a worldwide pandemic which would affect our societies for several years, causing the deaths of more than 6 million people in the world [19] and restrictions adopted to contain the virus causing permanent lifestyle changes [20].

1.1. First wave

February 2020 It did not take long before the first cases were detected in Italian citizens: on February 21, a 38-year-old man tested positive at Codogno hospital, and within a few hours, fourteen other people tested positive. On the same day, 200 km from Codogno, the first Italian victim of COVID-19 died. The man was a 78-year-old who lived in Vo' Euganeo, a city in the province of Padova. As a precautionary measure, the government suspended all public gatherings and non-essential commercial activities and closed workplaces, schools, and cultural places in ten municipalities in Lodi province, including Codogno and in Vo' Euganeo in Veneto. Furthermore, the government introduced the first restrictions on public transport, closing the train stations of Codogno, Maleo, and Casalpusterlengo [21].

As the epidemic diffused outside the hotspots areas, mainly in Lombardia, but also in Emilia-Romagna, Friuli-Venezia-Giulia, Veneto, Piemonte, and Liguria, restrictions were put in place in these Italian regions. These restrictions included school and museum closures, remote working, and the suspension of sports events. Moreover, all schools and universities on the national territory were closed.

March 2020 In early March, a new COVID-19 hotspot emerged in the Val Seriana area, affecting Alzano and Nembro municipalities significantly [22]. Despite this, the red zone was not extended to include the area. It remains a contentious decision with an ongoing investigation about its consequences on the area's severe COVID-19 impact in March 2020 [23]. Other research indicates that other regions, surrounding the town of Orzinuovi in the Brescia province and the city of Cremona, were also experiencing higher-than-average rates of positive cases during this period [24].

On March 8, Lombardia, along with some provinces of Emilia-Romagna, Veneto and Piemonte, was declared a red zone with maximum mobility restrictions. People were ordered to stay home; they could go out only for essential reasons like working or buying groceries and had to fill out a form to declare why they were leaving their households. The number of cases was rapidly increasing, particularly in the province of Bergamo. The restrictions were later extended to the entire national territory. Trenord responded to the new restrictions by reducing train rides by 8% in Lombardia, sanitising their trains, and advising passengers to wear masks and gloves and maintain social distancing [25].

During this period, Lombardia's railway mobility was significantly reduced, with an 85% decrease in daily passengers. Trenord cut its daily train rides by 40% to accommodate the new restrictions [26].

April 2020 On April 5, a significant milestone was reached as Italy reported a reduction in COVID-19 patients in intensive care for the first time since the pandemic began. This marked the beginning of the plateau phase of the epidemic.

After extending the lockdown period multiple times, citing its positive impact on slowing the spread of the virus [27], Italy's Prime Minister announced on April 26 that "phase two" of living with the virus would begin on May 4. This meant that four million Italians could return to work and people would be able to visit their relatives, signalling a slight easing of the stringent restrictions that had been in place for weeks.

May and June 2020 The month of May marked the end of the national lockdown and the beginning of a new phase in the fight against the virus. People gradually started to experience some of their old freedoms, starting on May 4.

Trenord announced the return of all suburban lines to 100% of their scheduled train rides, while regional lines increased from 40% to 60%. Travelling by train was heavily regulated: passengers had to wear masks and gloves, and only 50% of the seating capacity was available to maintain a one-meter distance between all passengers [28, 29].

Restrictions were loosened further on May 18, which marked the end of the first lockdown. Bars, restaurants, offices, and industries were allowed to reopen. People could move freely within their region (but not yet cross regional boundaries) without filling out any form. They were required to maintain a one-meter distance from each other and always wear masks.

On June 11, the beginning of "phase three" was announced, and on June 15, playgrounds, open-air cinemas, and theatres reopened, and people could travel freely throughout the nation.

1.2. Second wave

During the summer, all epidemic indicators reached their minimum levels since February. People resumed a life similar to before the pandemic, with only minor restrictions such as wearing a mask in public places.

On September 14, schools reopened after their national closure on March 4. Following the reopening of schools, Trenord decided that all its train lines (suburban and regional) would resume 100% of their scheduled rides [30]. The government also mandated that all public transportation operates at 80% of its total capacity, including seating and standing places [31].

Around the middle of September, the pharmaceutical giant Pfizer announced that they had successfully developed a COVID-19 vaccine that would be available on the market before the end of 2020. However, at the same time, the infection curve was constantly increasing all over Italy.

October 2020 By October 18, it was evident that Italy had entered the second wave of the pandemic, as the infection curve was growing exponentially.

On October 19, new restrictions were introduced to reduce infections. People were strongly advised not to invite more than six individuals into their homes at once, and bars and restaurants had to close in the evenings. Certain sports activities were limited, and celebrations were restricted to 30 people [32]. At first, no additional restrictions were placed on public transport. There was much concern about the potential for crowding on buses, metros, and trains during this period, and their role in spreading infections was discussed. Still, no definitive conclusions were reached [31].

On October 31, 31,758 daily new cases were reported, surpassing the maximum number of cases registered in a single day during the first wave, which was 6,557 on March 21. The situation was becoming increasingly critical, and new measures were necessary to curb the spread of the virus.

November 2020 On November 4, a new decree was introduced in Italy, implementing a three-tier system beginning on November 6. Italy was divided into yellow, orange, and red zones, corresponding to the epidemic's severity. The red zone experienced a lockdown with measures slightly less stringent than the first one. Additionally, a curfew from 10 p.m. to 5 a.m. was imposed in all zones. Lombardia was designated as a red zone with maximum restrictions from November 6 until November 29 [33, 34]. Due to the significant mobility restrictions similar to those of the March-April lockdown, the red zone period is often referred to as the second lockdown. Furthermore, public transport capacity is reduced to 50% in all zones (yellow, orange, and red) [35].

On November 13, the infection peak was reached, with over 40,902 new cases reported. After that, the red zones saw a decline in their epidemic indicators, leading to Lombardia being declared an orange zone starting on November 29.

December 2020 On December 13, Lombardia moved to the yellow zone with minor restrictions following further decreases in new infections. This marked the end of the second wave period. However, on December 18, a decree was issued stating that the holiday period would be subjected to red-zone restrictions throughout Italy to prevent a rise in infections following Christmas and New Year's Eve celebrations.

On December 21, the European Medicines Agency approved the vaccine for use in the European Union, and Italy administered the first vaccines on December 27.

1.3. An outlook on 2021

In 2021, most of the Italian population received the vaccine, while new variants of COVID-19 emerged, leading to further waves of infections in March and December. The three-tier system was expanded to include a white zone with few restrictions and then cancelled in the summer. After this, restrictions were mainly imposed on people who did not have the "Green pass," a certificate of vaccination or immunity due to a past infection.

2 | Literature Review

Since the COVID-19 outbreak at the beginning of 2020, multiple studies have tried to assess which relationship occurs between people's mobility and the spread of the disease [6, 7, 10–12, 36–51]. Several nations imposed lockdowns and preventive measures that significantly affected how people move. We can now evaluate the overall role of mobility in the evolution of the epidemic, together with the restrictions' impact, thanks to the willingness of public and private companies to supply data describing mobility.

2.1. COVID-19 impact on mobility

Researchers worldwide have tried to assess the impact of the disease spread and the non-pharmaceutical interventions (NPIs) put in place because of the virus [36–40].

In the United States, [36] quantitatively assesses the human mobility trend during COVID-19 to reveal how much policies impacted human movement, finding that government policies have little influence on it compared to the threat from the virus and the fear of being infected. Similarly, [37] combines GPS data about the average distance travelled by individuals at the country level with COVID-19 case data and estimates its impact on the reduction in mobility. They find that a rise in the local infection rate is associated with decreased mobility. Focusing on the Italian case, multiple studies were conducted: [38] analyses Facebook Italian mobility data, assessing how the 2020 lockdown affected the economy of municipalities, [39] considers the economic impact of the three-tier system adopted in Italy since November 2020 using mobile operator data, and [40] makes use of mobile positioning data to explain the impact of the restrictions on people's mobility and on the infection reproduction number R_t over time.

2.2. Mobility impact on COVID-19

After establishing the pandemic's strong influence on people's movement, the following question is to quantify the impact of mobility on the disease's spread. More specifically, we need to investigate which factors affected some epidemic indicators, such as the number of cumulated cases or the excess of mortality (i.e., the number of deaths from all causes during a crisis above and beyond what we would have expected to see under normal conditions [52]). A first hint of the spatial correlation of the pandemic in Italy is provided by [10], which analyses the excess mortality in Italy, identifying spatial clusters characterised by anomalous mortality compared to neighbouring territories. In the US, [41] finds a strong correlation between the changes in mobility patterns, measured by phone mobility data, and the decrease in COVID-19 growth rates, while [42] analyses the phenomenon of urban-to-rural migration in the US following lockdown announcements, showing that this migration possibly spread the disease faster in these areas.

[12] analyses data shared by mobile network operators to assess the mobility impact on the pandemic in France, Italy, and Spain. They consider the relationship between an epidemic indicator (excess deaths for France and Italy) and mobility data in the form of OD matrices provided by mobile operators. They choose to focus on mobility from a territory considered the outbreak of the epidemic (Haut-Rhin in France and Lodi province in Italy) to all the other departments or provinces in a few months of 2020 following the initial pandemic. They find that mobility alone can explain up to 92% of the epidemic variable associated with the initial spread of the disease in France and Italy. The findings confirm the significant impact of mobility on COVID-19 spread. Another notable result is the estimation of the lagged positive effect of reduced human mobility on the response variable, which is around 14-20 days.

[11] considers the Italian provinces and employs OD matrices again to construct functional curves describing mobility for each week of 2020 and 2021. A negative correlation is found between the mobility curves and curves expressing epidemic indicators in the form of mortality or infections. Similarly to the previous study, the temporal lag between the reduction in mobility and the actual mitigation of the epidemic is computed, leading to a result of 40 days (considering mortality concentration as the epidemic response). This study expands the findings of [12] since it investigates a vast territory (all of Italy) and a longer temporal span.

Other studies try to explain some data related to COVID-19 and the pandemic's impact on the economy, starting with mobility data. To cite only a few of them, [43] analyses the effects of mobility contraction on furloughed workers and excess deaths in Italy, [44] uses public mobility data provided by Google and Apple to show that there is an econometric causality between some mobility indicators and pandemic ones in Turkey, as well as [45] does with similar data for the whole Italian state, and [46] considering only Google mobility data for Lombardy.

After establishing that mobility and the epidemic phenomenon are strongly tied together, several studies apply different types of mobility data to infer COVID-19-related indicators. [47] considers data about the regional road infrastructure of Germany and identifies a negative association between the number of infected cases per capita and a metric measuring accessibility by road infrastructure between regions. This negative association is highly significant and robust to the addiction of potential confounding factors. Furthermore, the number of cases per capita is modelled considering the variation in mobility. A significantly positive effect of accessibility on mobility was found, while the mobility change negatively impacts the number of infected cases per capita.

[48] develops models to predict new cases of COVID-19 at county level in the US. The predictive features considered are based on spatiotemporal lags of infectious rates, human interactions, human mobility, and socioeconomic composition of counties. The final model selected is compared with the COVID-19 Forecast Hub Baseline model and improves its performance by 6.46% in the two-week and 20.22% in the four-week prediction horizon.

[49] is concerned with evaluating the set of restrictions to contain the spread of the disease. It aims to develop a model to generate forecasts of COVID-19 cases through publicly available data. The impact of NPIs is decomposed into two components: the behaviour change associated with the NPI and the resulting change in infections related to behaviour change. They develop two models: the behaviour model estimates how mobility in each country changes in association with the employment of NPIs, while the infection model considers the daily growth rate of infections as a function of human mobility. Considering the behaviour model, they found evidence that lockdown policies are associated with a substantial reduction in mobility, while considering the infection model, it was found that mobility data alone are sufficiently meaningful to forecast COVID-19 infections 7-10 days ahead at all geographic scales considered. Models excluding mobility data perform significantly worse, thus pointing to the leading role of mobility in forecasting.

[50] again employs mobility to model COVID-19 transmission to mainland China. A novel method for constructing asymmetric spatial weights based on population flow mobility data is proposed. These spatial weights are applied to a spatial econometric model, considering cumulative confirmed cases of COVID-19 as the response. The model significantly outperforms those using the traditional inverse distance weight matrix to describe spatial dependence.

This summary aims to show some examples of significant works in this field of research, considering studies employing statistical models. However, this does not exhaust the area: much work was done in modelling COVID-19 through mobility data, using methods such as compartmental models [53, 54], Hawkes process [55], neural networks [56], and much more.

2.3. The role of public transport

After the epidemic outbreak, much concern was raised about the role of public transport in the disease spread. In Italy, [5] reports that fear of pandemic diffusion on public transport led to a reduction of 50% in its usage in 2020, 42% in 2021 and 21% in 2022, compared to the 2019 levels.

The literature regarding this topic is narrow: [6] considers areas in England and Wales and finds that local areas with a significant fraction of people using public transport have more COVID-19 infections per 100,000 people. However, there is no consensus about public transport positively affecting the spread of the epidemic. In Italy, [51] analyses the relationship between the commuting network and the diffusion of COVID-19. This work does not explicitly relate to public transport. Still, it addresses mobility data from a new perspective compared to those presented previously in this chapter, considering the impact of the commuting phenomenon. Data is taken from the 2011 ISTAT census in the form of an OD matrix. Some metrics are defined, describing the commuting mobility phenomenon. They also include other variables potentially correlated with the excess of mortality (chosen as a target) and commuting patterns (such as but not limited to, population density, the proportion of males and elderly people in the municipality, and number of hospital beds). They build a fixed effects model and find that cities with larger shares of population commuting from and to their borders for labour motives tend to have higher excess mortality rates. However, [7] takes a step further, investigating the association between transit usage and the diffusion of COVID-19 in Italy. They estimate a fixed-effects and a quantile regression model, considering the excess of mortality as the target variable and an index expressing the share of the total population who moved daily by collective means of transport for labour or study. Again, they consider some confounding factors, as in [51]. The critical finding is that none of the coefficients associated with the variable expressing public transport usage is statistically significant. At the same time, those related to internal and external commuting flows are positively and strongly correlated with excess mortality.

2.4. Summary

Much work was done considering the association between COVID-19 and human mobility. Several studies show how COVID-19 and the restrictions adopted to limit it greatly influenced people's movement, but also that mobility is a crucial factor in explaining COVID-19-related variables: several kinds of models were built to explain and predict targets related to COVID-19, such as the number of new or cumulated cases, the number of per capita cases, or the excess of mortality in a specific territory considering various mobility-related regressors. There is evidence of the high importance of these features in predicting COVID-19. Another strand of research concerns the role of public transport: the real impact of the mode of transportation on the COVID-19 spread is highly unknown. Nevertheless, there is evidence in the literature that it may not have been as decisive as thought in the first period of the epidemic.

Our goal is to deepen the investigation of the role of mobility in the pandemic's spread, analysing an areal epidemic indicator in a much finer spatial granularity than the regional or provincial level often considered in the works presented in this Chapter. We will compare two spatial descriptions of two territories (first the entire Lombardia region and then a portion of it belonging to the provinces of Brescia, Bergamo and Milano), one based on mobility and the other on contiguity to assess the impact of mobility on the epidemic compared with a purely geographical description. While several works analysed spatial autocorrelation in an epidemic feature through contiguity-based spatial weights [10, 11], we found only another past work employing a mobility-based spatial description similar to ours [50]. In that case, the spatial weights were used to build a spatial econometric model for the diffusion of COVID-19 in China, while we will employ the mobility spatial description to investigate spatial autocorrelation in an areal epidemic feature.

Concerning the role of public transport in the disease's spread, we will focus on a specific kind of public transport, railway mobility. We will derive a dynamic representation of movements by train in a limited area of Lombardia through a limited portion of the Trenord railway network. These data will be employed to compare the spatial description derived from overall mobility (as described by the Regione Lombardia mobility dataset) with the dynamical one derived from public railway mobility and investigate spatial autocorrelation in the epidemic indicator. While we showed some studies relating to the public transport impact in this Chapter, we found no other works in past research investigating public transport mobility in the form of OD matrices relating to different transportation methods and their relationship with the pandemic.

3 | Datasets on mobility and epidemic

This study aims to analyse how mobility flows affected the spread of COVID-19 in 2020. Table 3.1 presents data used to reach this goal, divided into mobility and epidemic data, described respectively in Section 3.1 and Section 3.2.

Type of data	Description	Source
Epidemic	Death counts	ISTAT [9]
Mobility	OD matrix	Regione Lombardia [57]
Mobility	Ticket and counter data	Trenord

Table 3.1: Data summary

First, we consider mobility flows in an average workday of 2020, provided by Regione Lombardia (RL). These data cover the Lombardia area at a fine spatial granularity but do not represent actual movements in 2020 since they are derived from projections and rescaling of past data according to expected changes in mobility, like the opening of new highways. We present the RL dataset and report an exploratory analysis in subsection 3.1.1.

Because of the limited mobility description of the RL data, we introduce real movements only for a small area partially belonging to three provinces of Lombardia (Brescia, Bergamo and Milano) and construct real railway flows from the Trenord datasets presented in subsection 3.1.2. We postpone the exploratory analysis to Section 4.2 to better underline some key points after the presentation of the theoretical framework of the analysis.

This kind of data also enables us to face our second research question, assessing the impact of railway mobility on the epidemic compared to the overall mobility reported in the RL data.

Lastly, the ISTAT road distance dataset described in subsection 3.1.3 is needed to match every municipality to the closest station and assign the areal epidemic feature to the spatial granularity induced by stations.

Throughout the analysis, we will always model the epidemic response through death counts data provided by ISTAT [9], introduced in subsection 3.1.2. We now describe these datasets in detail.

3.1. Mobility data

3.1.1. Regione Lombardia Origin-Destination matrix

To model mobility in Lombardia, we use a dataset from Regione Lombardia describing people's average movements during a workday in 2020 [57]. These data were based on the ISTAT survey of 2011, together with contributions from public and private companies [58]. Movements are described in the form of OD matrices produced for the years 2014, 2016 and 2020. The OD matrices produced after 2014 have been derived, based on the 2014 matrix, from projected changes in transportation networks and rescaling accounting for changes in populations. Because of this, we will refer to this dataset with the term "projected" to underline that it is based on data collected prior to 2020.

This dataset describes expected movements and was released in 2019; thus, it does not account for the unexpected mobility changes caused by the COVID-19 pandemic in 2020. Consequently, we must be careful in using this dataset to represent 2020 mobility. While it is safe to assume that people moved as data portray in the first months of 2020, the pandemic outbreak at the end of February and the first restrictions put into place completely disrupted usual mobility trends, making the RL description likely unreliable.

The OD matrix estimates movements divided into origin, destination, time slot, reason and means of transport (considering the predominant means if multiple have been employed for the trip). The matrix refers to an average workday in Lombardia and considers 1525 areas (1450 of which internal to Lombardia, while the others represent other neighbouring Italian provinces, non-neighbouring Italian regions, neighbouring Swiss districts and foreign countries). Data consider 8 means of transport (car driver, car passenger, public transport by road, public transport by rail, motorcycle, bicycle, on foot and others) and 5 reasons to move (work, study, occasional, business, return). The space granularity mainly refers to municipalities, even if some small towns are aggregated, while some large cities are distributed over multiple areas. For more information about the distribution model applied by Regione Lombardia to construct the dataset, see [58].

Concerning the preprocessing operations applied to the dataset, we aggregate the time slots since we do not want to consider them for further analysis. We compute the sum of all the movements for every reason and means of transport and the sum considering only the railway public transport since we want to compare overall mobility with railway mobility in Chapter 5. We remove all the OD couples concerning areas outside of Lombardia and aggregate the zones referring to disaggregated large municipalities. The final OD matrix describes movements between 1360 zones referring to the 1504 municipalities in Lombardia.

Exploratory analysis of Regione Lombardia mobility data

The OD matrix of Regione Lombardia contains information about projected movements during an average workday in Lombardia for the year 2020. This matrix captures 9,229,377 trips taking place within Lombardia, with an average of 6,786 trips per zone (1360 zones) and 0.92 daily movements per person (given the population of Lombardia in 2020 was 10,027,602 [59]).

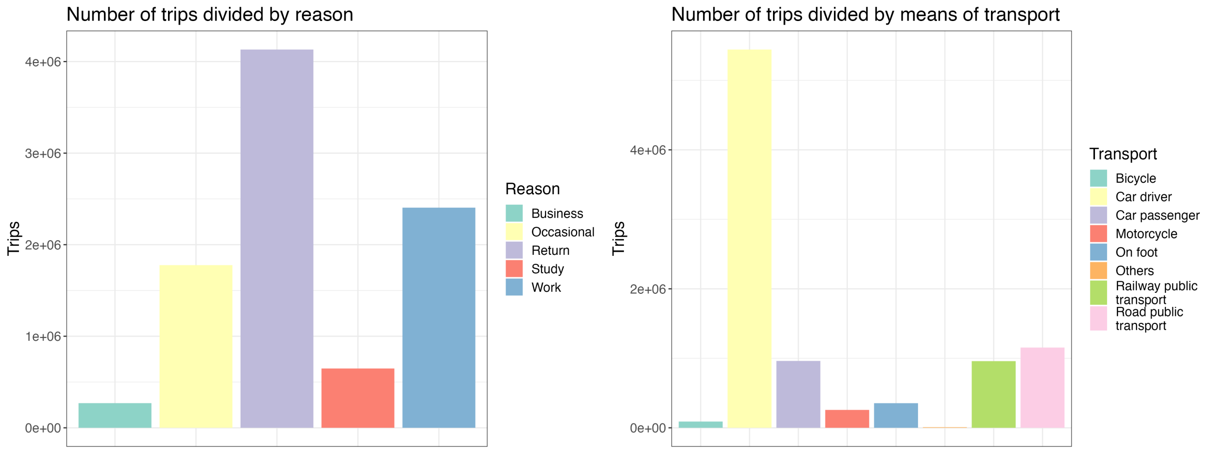
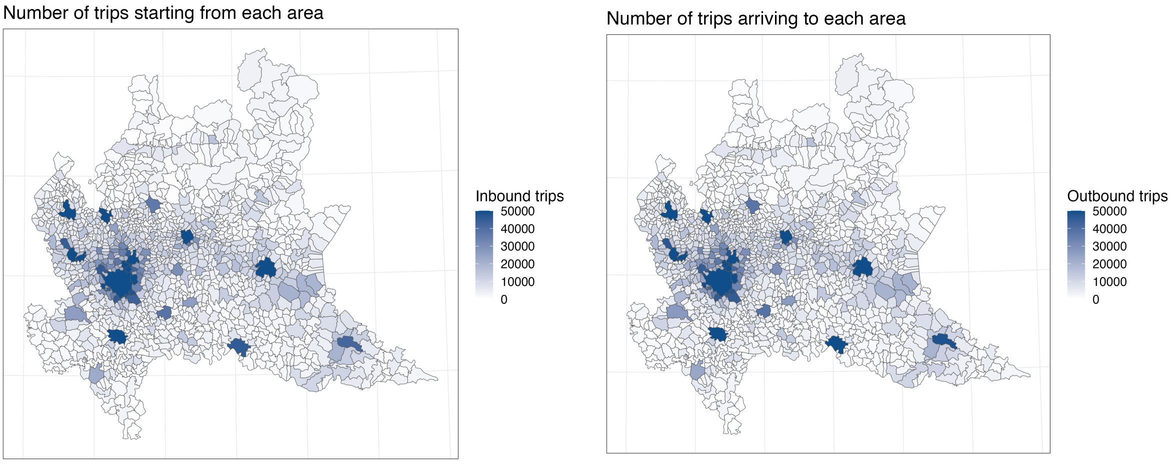


Figure 3.1: Trips distributions in the RL OD matrix

Figure 3.1 displays the distribution of trips by reason and transportation used. The principal reasons for travel are working or returning home, and most people use cars as their primary mode of transport. Railway public transport, which we will focus on in this work, is the fourth most used means of transport.

Figure 3.2 presents the number of inbound and outbound movements for each area, indicating that major cities experience the highest number of inbound and outbound trips due to their large populations. There are no visible differences between inbound and outbound trips: if there are many movements between areas i and j, there will also be many between j and i.

Lastly, Figure 3.3 summarises the mobility network at the province level, highlighting that most movements occur within the provinces. Apart from this, some territories are more strongly connected than others, such as Milano with Monza, Lodi, Pavia, Bergamo, Varese, and Bergamo with Brescia.



(a) Inbound movements

(b) Outbound movements

Figure 3.2: Inbound and outbound movements in the RL OD matrix

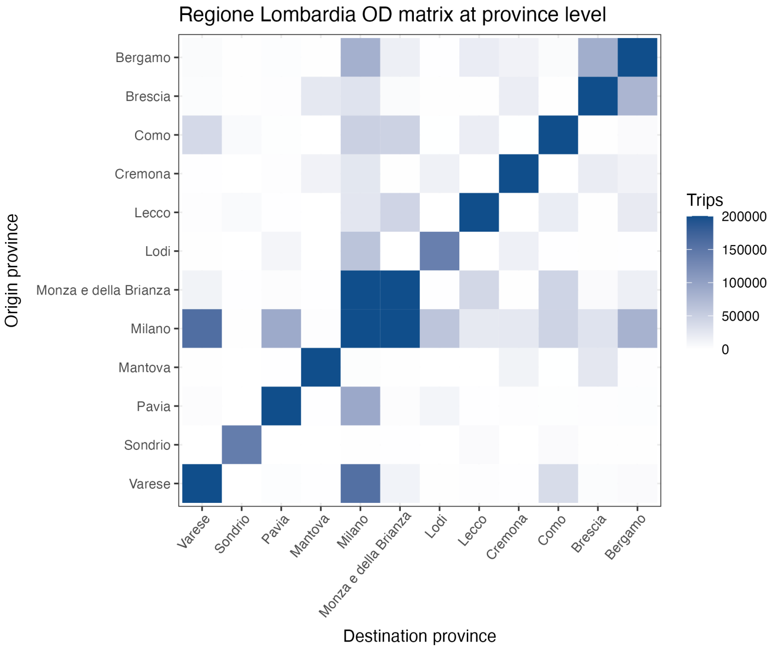


Figure 3.3: Movements in the RL OD matrix, divided by provinces

3.1.2. Trenord data

Data provided by Trenord aim to model people's movements by train during 8 months of 2020: from March to June and from September to December. We requested these months to compare two periods (Spring and Fall-Winter), each made by two months without restrictions due to COVID-19 and two months applying restrictions or lockdowns. We have two datasets to model the phenomenon, one concerning tickets and the other about flows of passengers boarding and dropping each train collected by the Automatic Passenger Counting (APC) system. Data involve the six train lines shown in yellow in Figure 3.4 travelling between Bergamo, Brescia and Milano provinces. These provinces were selected to compare territories where the epidemy spread at different speeds during 2020. Table 3.2 reports a summary of the six train lines involved in our study.

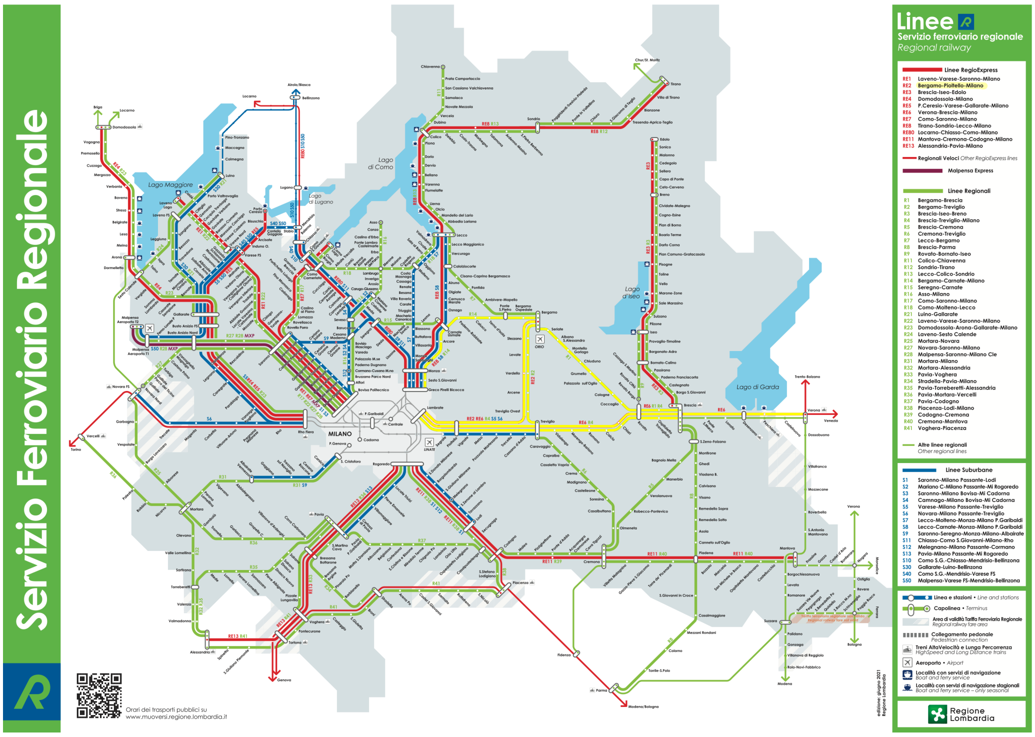

Servizio Ferroviario Regionale
Lago Maggiore
Lago di Como
Lago di Lugano
Lago d'Iseo
Lago di Garda
MILANO
Orari dei trasporti pubblici su www.muoversi.regione.lombardia.it
edizione: giugno 2021 Regione Lombardia
Linee
Servizio ferroviario regionale
Regional railway
Linee RegioExpress
RE1 Laveno-Varese-Saronno-Milano
RE2 Bergamo-Pioltello-Milano
RE3 Brescia-Iseo-Edolo
RE4 Domodossola-Milano
RE5 P.Ceresio-Varese-Gallarate-Milano
RE6 Verona-Brescia-Milano
RE7 Como-Saronno-Milano
RE8 Tirano-Sondrio-Lecco-Milano
RE80 Locarno-Chiasso-Como-Milano
RE11 Mantova-Cremona-Codogno-Milano
RE13 Alessandria-Pavia-Milano
Regionali Veloci Other RegioExpress lines
Malpensa Express
Linee Regionali
R1 Bergamo-Brescia
R2 Bergamo-Treviglio
R3 Brescia-Iseo-Breno
R4 Brescia-Treviglio-Milano
R5 Brescia-Cremona
R6 Cremona-Treviglio
R7 Lecco-Bergamo
R8 Brescia-Parma
R9 Rovato-Bornato-Iseo
R11 Colico-Chiavenna
R12 Sondrio-Tirano
R13 Lecco-Colico-Sondrio
R14 Bergamo-Carnate-Milano
R15 Seregno-Carnate
R16 Asso-Milano
R17 Como-Saronno-Milano
R18 Como-Molteno-Lecco
R21 Luino-Gallarate
R22 Laveno-Varese-Saronno-Milano
R23 Domodossola-Arona-Gallarate-Milano
R24 Laveno-Sesto Calende
R25 Mortara-Novara
R27 Novara-Saronno-Milano
R28 Malpensa-Saronno-Milano Cle
R31 Mortara-Milano
R32 Mortara-Alessandria
R33 Pavia-Voghera
R34 Stradella-Pavia-Milano
R35 Pavia-Torreberetti-Alessandria
R36 Pavia-Mortara-Vercelli
R37 Pavia-Codogno
R38 Piacenza-Lodi-Milano
R39 Codogno-Cremona
R40 Cremona-Mantova
R41 Voghera-Piacenza
Altre linee regionali Other regional lines
Linee Suburbane
S1 Saronno-Milano Passante-Lodi
S2 Mariano C-Milano Passante-Mi Rogoredo
S3 Saronno-Milano Bovisa-Mi Cadorna
S4 Camnago-Milano Bovisa-Mi Cadorna
S5 Varese-Milano Passante-Treviglio
S6 Novara-Milano Passante-Treviglio
S7 Lecco-Molteno-Monza-Milano P.Garibaldi
S8 Lecco-Carnate-Monza-Milano P.Garibaldi
S9 Saronno-Seregno-Monza-Milano-Albairate
S11 Chiasso-Como S.Giovanni-Milano-Rho
S12 Melegnano-Milano Passante-Cormano
S13 Pavia-Milano Passante-Mi Rogoredo
S10 Como S.G.-Chiasso-Mendrisio-Bellinzona
S30 Gallarate-Luino-Bellinzona
S40 Como S.G.-Mendrisio-Varese FS
S50 Malpensa-Varese FS-Mendrisio-Bellinzona
Linea e stazioni • Line and stations
Capolinea • Terminus
Area di validità Tariffa Ferroviaria Regionale Regional railway fare area
Collegamento pedonale Pedestrian connection
Treni AltaVelocità e Lunga Percorrenza HighSpeed and Long Distance trains
Aeroporto • Airport
Località con servizi di navigazione Boat and ferry service
Località con servizi di navigazione stagionali Boat and ferry service – only seasonal
Regione Lombardia

Figure 3.4: Trenord map of services [60]. Lines belonging to the study are coloured in yellow.

Line	Main stations
R1	Bergamo-Brescia
R2	Bergamo-Treviglio
R4	Brescia-Treviglio-Milano
R14	Bergamo-Carnate-Milano
RE2	Bergamo-Pioltello-Milano
RE6	Verona-Brescia-Milano

Table 3.2: Main stations of the Trenord lines involved in the study [61]

Ticket data

The ticket dataset reports tickets having origin or destination in one of the 46 stations belonging to one of the six train lines involved in the study. We have data for the 8 months cited above, plus two additional months (February and August 2020) where purchases for weekly or monthly subscriptions in March or September have been made. Table 3.3 reports a description of the dataset's variables.

Variable	Description
`Destination code`	Code of the origin train station (e.g., S01529)
`Origin code`	Code of the destination train station
`Date of emission`	format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19)
`Product Type`	Ticket types are "Annual subscription", "Monthly subscription", "Weekly subscription", "Other operator", "Ordinary ticket", "Carnet", "Supplementary correction", "Special rates initiative", "Malpensa Express", "Unpaid remains", "Sanctions", "Supplements"
`Quantity`	

Table 3.3: Summary of variables in the Trenord ticket dataset

There are some missing ticket data to deal with in our analysis concerning two critical cases:

- Tickets concerning couples of stations both internal to the Milan municipality do not appear in the data, while we have data about tickets internal to other cities like Bergamo or Treviglio.

- Tickets with origin or destination in station Verona Porta Nuova are sold applying an interregional rate that does not appear in Trenord data. Thus, we have no information about tickets for passengers boarding or dropping in Verona.

Counter data

The counter dataset reports passengers boarded and dropped at each station for each train ride belonging to one of the six train lines of the study during the 8 months. Data are recorded through the APC system, installed on approximately 40% of the Trenord fleet in 2021 [62].

Systems to effectively count the number of passengers on railway networks are crucial to infer the characteristics of travel demand and to optimise the service, and their installation on public services in recent times has been increasing [63]. APC systems count passengers boarding and dropping the train through sensors placed on the train doors. There is an error in accuracy in the passengers counting procedure, usually ranging from 5% to 10%. To provide reliable estimates, the train must be equipped with sensors at every train door. When the train is partially equipped (not less than 3/8 of the composition) by APC, estimates are made. If this fails, estimates can be made by interpolations considering the same train on the same day of the previous week for all the last four weeks. If this estimate is not accurate enough, an average of the whole previous week's working days is computed, then using the last 30 days, and so on.

The dataset reports a status describing the outcome of the counting process:

- "-5": Ride with neither counters nor travel data. The ride has remained in status "Scheduled".
- "-4": Ride cancelled at the end of the mission.
- "-3": Ride cancelled at the beginning of the mission.
- "-2": Ride completely cancelled.
- "-1": Ride with invalid data (nor APC system nor interpolations).
- "0": Ride wholly equipped with APC.
- "1": Ride not equipped with APC, whose passenger counts are estimated via interpolations.
- "2": Ride partially equipped (3/8 of the composition has the APC system).
- "3": Ride estimated due to on-board anomalies, ride wholly equipped with APC.

- "4": Ride estimated due to on-board anomalies, ride partially equipped with APC.
- "5": No counters data received.
- "6": Ride with incorrect counters data.

States $[0, 1, 2, 3, 4]$ provide valid data through equipped or partially equipped train rides or interpolations. States $[-5, -4, -3, -2]$ correspond to cancelled train rides and states $[-1, 5, 6]$ correspond to missing data.

Table 3.4 reports a description of the dataset's variables.

Variable	Description
`Date`	Scheduled departure date, format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19)
`Day of the week`	"0" - Monday, "1" - Tuesday, "2" - Wednesday, "3" - Thursday, "4" - Friday, "5" - Saturday, "6" - Sunday and holidays, "-1" - race with invalid data (i.e., with status different from Completed or Recalculated)
`Train code`	Train ride code (e.g., 10913)
`Train direction code`	Code describing the direction of the train on the line (e.g., D017)
`Line`	Code of the railway line (e.g., R1)
`Departure station`	Code of the departure station (e.g., S01529)
`Arrival station`	Code of the arrival station (e.g., S01529)
`Scheduled departure time`	Departure time of the train, format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19)
`Scheduled arrival time`	Arrival time of the train, format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19)
`Progressive index of the stop station`	Index describing the progressive order of stops in the line
`Stop station code`	Code of the stop station (e.g., S01529)
`Station entry time`	Date of arrival at the stop station, format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19). The field is empty if the stop station code coincides with the departure station code
`Station entry delay`	Delay between scheduled and actual station entry time, expressed in minutes
`Station exit time`	Date of departure from the stop station, format "MM/DD/YYYY hh:mm:ss" (e.g., 02/09/2020 11:51:19). The field is empty if the stop station code coincides with the arrival station code

Variable	Description
`Station exit delay`	Delay between scheduled and actual station exit time, expressed in minutes
`Number of boarded passengers`	
`Number of dropped passengers`	
`Number of passengers on board`	
`Maximum number of registered passengers on board`	
`Cancelled station`	True or False
`Fleet code`	“0” - TSR, "1" - TAF, "2" - VIVALTO, "3" - MD, "4" - PR, "5" - DP, "6" - CoradiaAeroportuali, "7" - CoradiaRegionali, "8" - GTW, "20" - Caravaggio, "99" - Generic (Not specified fleet)
`Number of train cars`	(Net of any Locomotive and GTW traction unit, cars for which the number of seats offered is 0)
`Total train capacity`	
`APC system capacity`	
`UDT list`	
`APC system status`	Status describing the outcome of the APC counting process. Values are "-5", "-4", "-3", "-2", "-1", "0", "1", "2", "3", "4", "5", "6"

Table 3.4: Summary of variables in the Trenord counter dataset

3.1.3. ISTAT road distance data

The ISTAT road distance dataset [64] is essential for our analysis to build a connection between municipalities and stations, defining which cities refer to each station. This dataset reports the road distances in meters and the drive times (in minutes) between every couple of Italian municipalities. The distances were computed from geographical computer systems and a commercial road graph. Every path’s length was calculated from the city hall of the origin municipality to the city hall of the destination, considering ideal traffic conditions and average travel speed for every arc constituting the road path.

3.2. Epidemic data

As epidemiological source, we make use of the number of death counts from all causes of people aged 70 or more years to model the epidemiological outcome. We decided to focus on death counts for the reasons explained in [10].

Reports of the mortality data for years from 2011 to 2021 can be found in [9]. Moreover, we considered only mortality of people aged 70 years or older since [10] shows that this age class suffers from the greatest mortality risk when infected and can be used to model spatial patterns and pandemic effects. We select only mortality data for people over 70 years old residing in Lombardy in the ISTAT dataset and aggregate it over a weekly basis.

3.3. Analysis summary

In the upcoming Chapters, we will develop and apply the statistical methodologies needed to analyse the datasets described in this Chapter.

Firstly, we will define a pipeline to estimate dynamic OD matrices describing movements by train in a limited area consisting of six train lines of the Trenord railway network in 8 months of 2020. To achieve this, we will employ the ticket and counter datasets provided by Trenord, described in subsection 3.1.2 and apply the Furness method for trip distribution modelling which we will present in Chapter 4, together with the pipeline we developed to address the estimation problem.

In Chapter 5, we will analyse the spatial autocorrelation of mortality rates (derived from ISTAT death data described in section 3.2) through spatial descriptions defined from mobility data. In order to achieve this, we will apply some methods of spatial data analysis. This will include defining contiguity and mobility-based spatial weights, testing for positive spatial autocorrelation through global Moran indexes, and assessing local spatial autocorrelation by computing local Moran indexes. Initially, we will examine the entire Lombardia region and investigate spatial autocorrelation in the mortality rates within municipalities' spatial granularity. To accomplish this, we will utilise the mobility data released by Regione Lombardia presented in subsection 3.1.1. However, it is essential to note that this dataset does not provide a reliable mobility description after the pandemic outbreak at the end of February, as it is not estimated from data collected in 2020, but derived from past data according to projections. Thus, we will supplement this static mobility description with the dynamic Trenord OD matrices derived in Chapter 4.

Thanks to these dynamic data, we will repeat the spatial analysis in a limited area of

Lombardia and spatial granularity defined by assigning every municipality to the closest train station. We will use the ISTAT road distance dataset described in subsection 3.1.3 to match municipalities with the closest station. The OD matrices derived by Trenord data will allow us to have a dynamic reliable description of mobility trends throughout 2020 and to assess in detail the role of a specific kind of public transport (public railway transport) in the epidemic spread.

All of the analyses in this work are be carried out using the `R` software [65]. In particular, we used package `mipfp` [66] to perform the Furness method and package `spdep` [67] for spatial analysis. Regarding mortality data, we used some code developed by [10] for preprocessing operations.

The code to reproduce the analyses presented in this thesis is available at `https://github.com/GretaGalliani/spatial_analysis_mobility_pandemic`.

4 | Estimation of Trenord Origin-Destination Matrices

This Chapter describes the procedure to estimate OD matrices representing movements by train in 8 months of 2020 through six train lines of the Trenord network.

More in detail, Section 4.1 recalls some theoretical notions about the Furness method in trip distribution modelling and presents the four-step pipeline we developed to estimate the OD matrices from the ticket and counter data described in Chapter 3. Section 4.2 explores the available data about railway mobility in the form of ticket and counter datasets and highlights some criticalities. Section 4.3 presents the results obtained by applying the pipeline developed to transform the initial data into 37 weekly OD matrices. Then, Section 4.4 builds a spatial granularity matching municipalities with Trenord stations and compares the RL and Trenord mobility data, assessing their relationship through 2020. Finally, Section 4.5 discusses the results obtained and the criticalities in the estimation procedure.

The estimated OD matrices represent actual railway mobility in 2020, as opposed to the projected mobility (estimated from data prior to 2020) portrayed by the RL dataset. Both the projected static RL OD matrix and the estimated dynamical Trenord OD matrices derived in this Chapter will be used in the next Chapter to perform a spatial analysis of the relationship between an epidemic indicator and mobility flows.

4.1. Methodology

In this Section, we present the theoretical properties of the Furness method of trip distribution modelling. Trip distribution modelling predicts the number of trips between origins and destinations in a transportation network, and the Furness method [68], also known as Iterative Proportional Fitting [69], is a technique used in this framework. It is a widely used approach in transportation planning, and it has been applied in various studies (such as [70, 71]) to develop transportation plans and evaluate the impact of transportation infrastructure projects. In our work, we employ the Furness method to estimate actual trips by train that occurred in 8 months of 2020 in a limited area of Lombardia. The results, in the form of dynamic OD matrices, will provide us with data accurately describing weekly railway mobility in the area. The purpose of these dynamic data is twofold: they describe actual mobility trends through 2020 and allow us to address the role of a specific kind of public transport believed to be a primary carrier in epidemic diffusion. The next Chapter will use this data to derive a spatial description based on mobility flows and to assess the relationship between mortality rates and this spatial description. The Trenord dynamic OD matrices provide us with the only reliable mobility representation following the epidemic outbreak, as the RL data can not be trusted to describe real mobility flows after the end of February.

We now present the theoretical properties of the Furness method and then detail the four-step pipeline we developed to derive the weekly OD matrices.

4.1.1. Furness method

We aim to estimate the number of trips for each couple of nodes likely to be made per time unit. Fixing the time unit, we start with an origin and destination survey producing a seed matrix. In our case, we derive the seed matrix from ticket data. This matrix has the form

$$\begin{array}{c} \begin{bmatrix} t^*_{11} & \cdots & t^*_{1J} \\ \vdots & & \vdots \\ t^*_{I1} & \cdots & t^*_{IJ} \end{bmatrix} \begin{matrix} q_1 \\ \vdots \\ q_I \end{matrix} \\ \begin{matrix} b_1 & \cdots & b_J & \quad u \end{matrix} \end{array}$$

where t^*_{ij} is the number of trips beginning in zone i and ending in zone j, I is the number of zones where the trip can begin and J is the number of zones where the trip can end, $q_i = \sum_{j=1}^{J} t^*_{ij}$ is the number of trips beginning in zone i, $b_j = \sum_{i=1}^{I} t^*_{ij}$ is the number of trips ending at zone j and u is the total number of trips. Since this matrix is generated through a survey, its row and column totals q_i and b_j do not generally equal the estimates of the trips starting and ending in each zone. Let estimates of the actual number of trips beginning in each zone be $p_1, \dots, p_I$ and let $a_1, \dots, a_J$ be the estimates for the number of trips ending in each zone. In our application, we derive estimates of real trips from counter data. The total number of trips beginning must equal the number of trips ending such that

$$\sum_{i=1}^{I} p_i = \sum_{j=1}^{J} a_j = v \tag{4.1}$$

The trip distribution problem is to derive from matrix t^*_{ij} a forecasting matrix t_{ij} whose row and column totals are respectively $p_1, \dots, p_I$ and $a_1, \dots, a_J$:

$$\begin{array}{ccc}\begin{bmatrix} t_{11} & \cdots & t_{1J} \\ \vdots & & \vdots \\ t_{I1} & \cdots & t_{IJ} \end{bmatrix} & \begin{array}{c} p_1 \\ \vdots \\ p_I \end{array} \\ \begin{array}{ccc} a_1 & \cdots & a_J \end{array} & v \end{array}$$

To derive matrix t_{ij}, we need to iteratively find constants by which to multiply the elements of the original matrix t^*_{ij}.

Furness method provides an answer to the trip distribution problem. At each passage, a matrix $t^{(n)}_{ij}$ is obtained by multiplying the previous matrix $t^{(n-1)}_{ij}$ by a suitable constant $x^{(n)}_{ij}$. Thus

$$\begin{aligned} t^{(1)}_{ij} &= x^{(1)}_{ij} t_{ij} \\ t^{(n)}_{ij} &= x^{(n)}_{ij} t^{(n-1)}_{ij} \quad \text{for } n \geq 1 \end{aligned}$$

$t_{ij} = \lim_{n\to\infty} t^{(n)}_{ij}$ is the limiting matrix and $\lim_{n\to\infty} \prod_{k=1}^{n} x^{(k)}_{ij}$ is the required multiplying factor for t^*_{ij}.

We now report some theoretical results, whose proofs can be found in [68].

Theorem 4.1. *Given*

1. *Any $I \times J$ matrix T^*, all of whose elements $t^*_{ij} > 0\ \forall i = 1, \dots, I\ \forall j = 1, \dots, J$*
2. *Any set of numbers $p_1, \dots, p_I$ with $p_i > 0\ \forall i = 1, \dots, I$, and $\sum_{i=1}^{I} p_i = v$*
3. *Any set of numbers $a_1, \dots, a_J$ with $a_j > 0\ \forall j = 1, \dots, J$, and $\sum_{j=1}^{J} a_j = v$*

Then there exists at most one $I \times J$ matrix T whose elements t_{ij} satisfy:

$$\sum_{j=1}^{J} t_{ij} = p_i \quad \forall i = 1, \dots, I \tag{4.2}$$

$$\sum_{i=1}^{I} t_{ij} = a_j \quad \forall j = 1, \dots, J \tag{4.3}$$

and

$$\begin{gathered} t_{ij} = r_i s_j t^*_{ij} \textit{ for some positive numbers } r_1, \dots, r_I \textit{ and } s_1, \dots, s_J \textit{ and} \\ \forall i = 1, \dots, I\ \forall j = 1, \dots, J \end{gathered} \tag{4.4}$$

Theorem 4.2. *Consider a matrix T^*, numbers $p_1, \dots, p_I$, and $a_1, \dots, a_J$ as in Theorem 4.1. Define a sequence of matrices $t_{ij}^{(1)}, t_{ij}^{(2)}, \dots$, as follows (this is the Furness method of iteration):*

$$\begin{aligned} t_{ij}^{(1)} &= \frac{p_i}{\sum_{k=1}^{J} t_{ik}^*} \\ t_{ij}^{(2n)} &= \frac{a_j}{\sum_{k=1}^{I} t_{kj}^{(2n-1)}} t_{ij}^{(2n-1)} \quad \text{for } n \geq 1 \\ t_{ij}^{(2n+1)} &= \frac{p_i}{\sum_{k=1}^{I} t_{ik}^{(2n)}} t_{ij}^{(2n)} \quad \text{for } n \geq 1 \end{aligned} \tag{4.5}$$

Then $t_{ij}^{(m)}$ tends to a limit as $m \to \infty$. Let $\lim_{m\to\infty} t_{ij}^{(m)} = t_{ij}^0$. Then $t_{ij}^0 = t_{ij}$, where t_{ij} is the unique matrix defined by Equations (4.2) to (4.4) of Theorem 4.1. Hence the matrix t_{ij} always exists.

These two theorems are based on the assumption that the original matrix T^* elements are such that $t_{ij}^* > 0 \ \forall i \in 1, \dots, I \ \forall j \in 1, \dots, J$, which is a sufficient condition for the existence of matrix T. However, we should take particular care when dealing with zero cells as Furness method performs no adjustments for these entries: if $t_{ij}^* = 0$, then $t_{ij} = 0$. Zero cells are thus interpreted as an impossibility of travel between the two zones. Another assumption is $p_i > 0 \ \forall i \in 1, \dots, I$ and $a_j > 0 \ \forall j \in 1, \dots, J$. However, this assumption causes no particular problems when removed since, in this case, the corresponding row or column of matrix T will be empty, and Equation (4.5) will not be applied. Consequently, no journeys would be expected to begin or end at that particular zone.

Since Furness method is an iterative method, an algorithm performs Equation (4.5) of Theorem 4.2 iteratively until a number of maximum iterations is achieved or the following stopping criterion is reached:

$$\max \left| t_{ij}^{(n-1)} - t_{ij}^{(n)} \right| < tol \quad \forall i = 1, \dots, I \ \forall j = 1, \dots, J$$

Before applying the Furness algorithm, [72] proposes three statistics to perform a test of coherency between the seed matrix T^* and the set of marginals $\mathcal{M} = \{p_i, a_j \ \forall i = 1, \dots, I \ \forall j = 1, \dots, J\}$, whose null hypothesis can be informally expressed as

$$H_0 : \text{data } T^* \text{ agree with } \mathcal{M} \qquad \text{vs} \qquad H_1 : \text{data } T^* \text{ do not agree with } \mathcal{M}$$

We apply the test based on the Wald statistic in our applications since it is the only statistic able to handle matrices T^* containing zero cells.

After computing T, we can measure the concordance between the final matrix and the marginals $p_1, \dots, p_I$ and $a_1, \dots, a_J$ calculating the margins error as the maximum deviation between each generated and desired margins as

$$\begin{aligned} \epsilon_{row} &= \max_i \left| p_i - \sum_{j=1}^{J} t_{ij} \right| \\ \epsilon_{col} &= \max_j \left| a_j - \sum_{i=1}^{I} t_{ij} \right| \end{aligned} \tag{4.6}$$

Critical issues

Some literature exists on critical issues of the Furness method [73]. We describe the main problems encountered in our application, proposing possible solutions.

Zero cell problem A consequence of the Furness method definition presented above is that a zero cell in the seed matrix t^*_{ij} will also be zero in the limit matrix t_{ij}. However, sometimes a zero value in the seed matrix is not associated with the impossibility of travel between the two zones: a zero can also result from an incorrect estimation during the survey. Therefore, some of these cells could not be actual zeroes. We should identify and correct these entries. [73] proposes two approaches when dealing with this problem:

1. Circumventing the occurrence of zeroes values
2. Replacing zeroes with suitably small values

The first approach is achieved via category aggregation or skipping the task of fitting tables and applying a simulation-based method. We use the second approach, replacing zero cells with a 1. However, it has been shown that this may severely bias zone tables and does not necessarily improve the fit. Because of this, we apply an original approach, performing a posteriori binomial test on the modified entry, testing the hypothesis

$$H_0 : \frac{t_{ij}}{p_i} > \frac{1}{p_i} \qquad \text{vs} \qquad H_1 : \frac{t_{ij}}{p_i} \leq \frac{1}{p_i} \tag{4.7}$$

We aim to test if the a posteriori value t_{ij} is significantly smaller than the a priori artificial value $t^*_{ij} = 1$. If we have statistical evidence that the value t_{ij} is significantly smaller than 1, we report the cell value to zero and apply Furness again.

Since this binomial test has to be run for every cell originally zero in t^*_{ij}, we apply the False Discovery Rate controlling procedure [74] to reduce the number of false discoveries. The level of significance is set at 0.05.

Marginals consistency Another assumption is expressed in Equation (4.1), which states that the total number of beginning trips should equal the total number of ending trips. However, this assumption is violated in our application since marginals $p_1, \ldots, p_I$ and $a_1, \ldots, a_J$ suffer from estimation errors. [66] provides a solution to overcome the consistency assumption: when consistency is violated, we shift to probabilities, defining

$$\begin{aligned} \pi^*_{ij} &= \frac{t^*_{ij}}{\sum_{ij} t^*_{ij}} \\ \rho_i &= \frac{p_i}{\sum_i p_i} \\ \alpha_j &= \frac{a_j}{\sum_j a_j} \end{aligned} \tag{4.8}$$

This operation recovers consistency since $\sum_{i=1}^{I} \rho_i = \sum_{j=1}^{J} \alpha_j = 1$. Then, Furness method is performed, using π^*_{ij} as seed matrix and $\rho_1, \ldots, \rho_I$ and $\alpha_1, \ldots, \alpha_J$ as marginals. The result is matrix π_{ij} with $\sum_{ij} \pi_{ij} = 1$. Every cell defines the probability of a single trip occurring from zone i to zone j. To recover the OD matrix estimating the number of actual trips between each couple of zones, we have to multiply matrix π_{ij} for the actual number of total trips v. We can choose v either as the total number of beginning trips $\sum_{i=1}^{I} p_i$ or the total number of ending trips $\sum_{j=1}^{J} a_j$.

Integer conversion The matrix t_{ij} entries have a natural interpretation as the number of trips starting from zone i and ending in zone j. As such, the cells should be integer numbers, while Furness method outputs a matrix containing non-integer values. We decide to apply integer conversion to the output matrix, aware that this approach has disadvantages. Some of them are pointed out by [73], which discusses the drawbacks of integer conversion regarding information discrimination: cells containing values of 0.501 are treated the same as cells containing 0.999 after this operation. Another question is whether the seed t^*_{ij} and marginals p_i and a_j should be integers. These values also have a natural interpretation in terms of estimated trips for t^*_{ij} and number of trips beginning in zone i for p_i and ending in zone j for a_j. The rounding choices will be described more carefully in the next subsection since they are the natural consequence of the procedures used to construct these quantities.

4.1.2. Pipeline to estimate Trenord Origin-Destination matrices

The theoretical framework described in subsection 4.1.1 will be employed to derive 37 weekly OD matrices describing movements by train through six train lines (connecting 46 stations) of the Trenord network in 8 months of 2020. The pipeline we built to reach this goal takes in input the ticket and counter datasets described in subsection 3.1.2 and develops in the four steps shown in Figure 4.1.

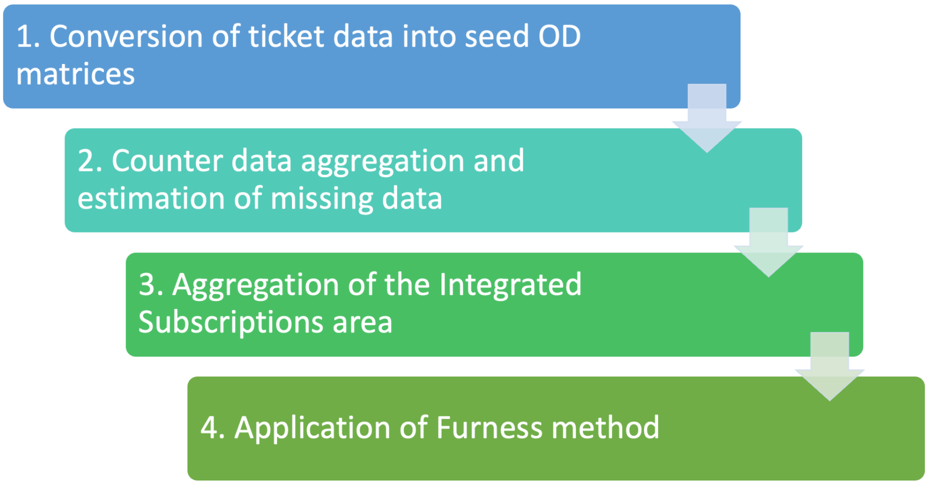

1. Conversion of ticket data into seed OD matrices
2. Counter data aggregation and estimation of missing data
3. Aggregation of the Integrated Subscriptions area
4. Application of Furness method

Figure 4.1: Pipeline to estimate the Trenord weekly OD matrices

This Section details every step of the procedure, highlighting the main issues encountered and the solutions adopted. It is important to note that this process could be adapted to other types of transportation networks, provided that we could derive a seed matrix from ticket data, paid tolls or turnstile counts, and margin vectors from estimates of trips beginning and ending in each zone.

Our pipeline introduces some novel features that differentiate it from previous work in this field, whose contribution to our estimation process will be discussed in Section 4.5:

1. Our procedure converts ticket purchases into estimated trips with an origin and a destination assigned to specific periods (weeks). The assumptions used in this framework have been discussed with the data provider. Still, a sensitivity analysis could be conducted in the future to assess the robustness of the final OD matrices to the ticket assumptions.

2. We developed a model to estimate the number of passengers boarded and dropped at each station in each week of the study, accounting for missing counter data.

3. We apply a correction of zero values in the seed matrices, which enables us to allocate movements for OD paths where the ticket allocation process does not place any trips. This procedure is essential to overcome the limitations of the translation of tickets into estimated trips since we are dealing with significant uncertainties in

allocating trips in each time period.

Data preparation

To prepare data for the application of the pipeline, we apply a time aggregation since assumptions are needed to convert ticket data into movements, and we deemed daily granularity too fine for our purposes. We aggregate our data into weeks starting Monday and ending Sunday. Since our study includes eight months of 2020, as stated in subsection 3.1.2, we build 37 weekly OD matrices in the next steps of the pipeline. Four of them (weeks 08, 26, 35 and 52 of the year, following the numbering according to which the first day of the year belongs to week 00) contain data for a fraction of the seven days, namely 1, 2, 6 and 4 days. Figure 4.2 displays the correspondences between the 2020 calendar and the weeks' numbers $w \in W$, highlighting the weeks belonging to the study period.

Considering tickets, we have data concerning February and August, where weekly and monthly subscriptions for the first weeks and months of March and September have been bought. Attribute `Quantity` in the ticket dataset can assume fractional values since rates corresponding to tickets combined with ATM (the society responsible for public transport in the Milan area) are reported as 0.5. We have no means of discovering how many tickets of this kind were sold, so we resolve to ceil attribute `Quantity`.

Conversion of ticket data into seed OD matrices

First, we have to convert every record in the ticket data into one or more estimated trips between stations, one being the origin and the other the destination. [75] provides rules for each ticket and subscription. We introduced a list of assumptions about the relationship between each record and estimated trips:

- Ordinary tickets, special rates initiatives, and additional exactions are translated as 0.5 trips between origin and destination and 0.5 between destination and origin by the end of the corresponding week. This is because each ticket can be used in either direction between the two stations for which it has been emitted, so we split the number of trips evenly in the two directions.

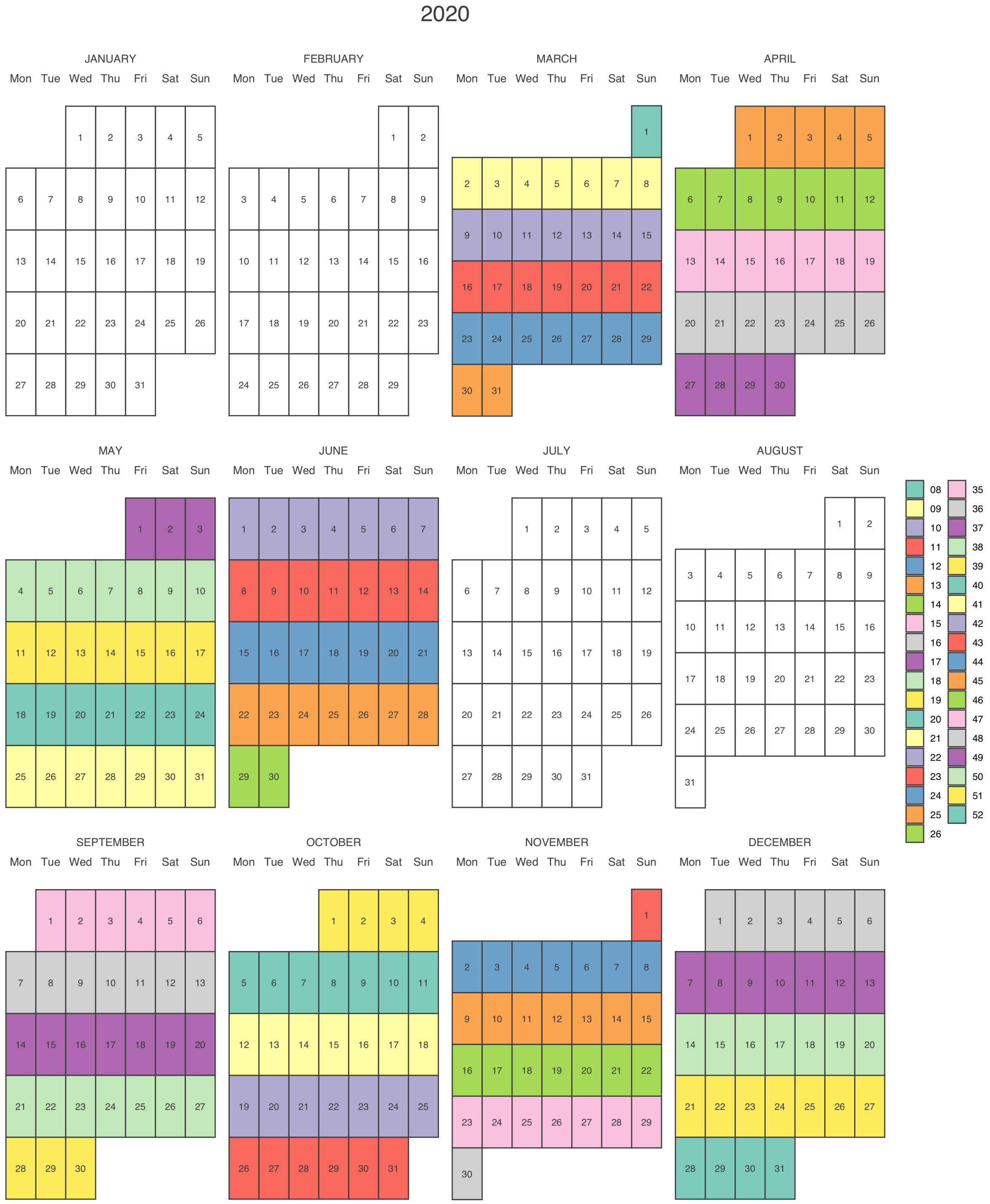


Figure 4.2: Weeks of the study in the 2020 calendar

- For carnets, we randomly extract 5 days in the 30 days following the carnet's purchase. We suppose a round trip is made in the 5 days drawn and then aggregate weekly. We chose the period of 30 days to extract the trips because it was previously the validity period of the carnet, while now carnets do not have an expiration date.

- For weekly subscriptions, we suppose 5 round trips attributed to the current week if the subscription is bought between Monday and Wednesday, to the following week if the subscription is purchased between Thursday and Sunday.

- For monthly subscriptions, round trips are distributed into the month's weeks starting from the day of selling. The month of usage is the current month if the subscription is bought before the 22^{nd} of the month or the following month if it is purchased on the 22^{nd} or the days after. We suppose 5 round trips for full weeks (i.e., entirely belonging to the subscription month). For partial weeks, we use the correspondences specified in Table 4.1 by computing and rounding the proportion $\frac{5\ round\ trips}{7\ days} * n\ partial\ days$. For instance, if a monthly subscription was bought on May 31 for June 2020, this subscription would be translated as 5 round trips for weeks number 22, 23, 24 and 25 and one round trip for partial week 26 since only two days of week 26 (from June 29 to July 5) belong to the month of June, after which the monthly subscription expires.

Number of partial days	Number of round trips
1	0
2	1
3	2
4	2
5	3
6	4

Table 4.1: Correspondences between partial week days and the number of round trips estimated in the ticket OD matrix for monthly and yearly subscriptions

- For yearly subscriptions, we attribute 5 round trips to each complete week starting from the day of purchasing and ending the last day of the 12^{th} month after purchase, applying the same convention to uncomplete weeks (if any) described in Table 4.1.

At the end of the process, the last passage is to put negative quantities to 0 as negative amounts describe transfers and when transfers are more than the actual tickets, we assume that the number of tickets estimated trips between those stations is 0.

Applying these assumptions, we obtain 37 weekly OD matrices $T^{*[w]}$ describing the ticket-estimated trips between each station. Because there is no distinction between ticket direction (a ticket bought between stations i and j can be used both to go from i to j and from j to i), the matrices are symmetrical. The matrices contain fractional values, but we do not round them since these values derive from the probability interpretation to attribute single trips to the origin-destination path or to the destination-origin one with a probability of 0.5.

Counter data aggregation and estimation of missing counter data

The second step of the estimation pipeline is to aggregate counter data and build margin vectors describing the total number of boarded and dropped passengers for each couple (i, w) of station $i \in S$ and week $w \in W$.

As explained in subsection 3.1.2, the APC system reports a status describing the outcome of the counting process for each train ride. States $[0, 1, 2, 3, 4]$ provide valid data through equipped or partially equipped train rides or interpolations. States $[-5, -4, -3, -2]$ correspond to cancelled train rides and states $[-1, 5, 6]$ correspond to missing data. These missing data must be estimated to model the number of people travelling weekly from and to station i.

First, we want to consider valid data (i.e., data from rides with APC states in $[0, 1, 2, 3, 4]$). For each couple of station and week (i, w), we compute the partial number of boarded and dropped passengers $boarded\_partial_i^{[w]}$ and $dropped\_partial_i^{[w]}$, summing boarded and dropped passengers considering valid train rides data registered in station i in the days of week w. We also compute an index expressing how much the couple is covered by counter data, defined as

$$coverage_i^{[w]} = \frac{\#\{train\ rides\ having\ valid\ counters\ data\}_i^{[w]}}{\#\{total\ train\ rides\}_i^{[w]}} \tag{4.9}$$

$coverage_i^{[w]}$ ranges in the interval $[0, 1]$, with $coverage_i^{[w]} = 1$ indicating that every train for the couple (i, w) has valid counter data, either from train rides wholly equipped with APC or from interpolations. In contrast, $coverage_i^{[w]} = 0$ corresponds to no valid counter data for couple (i, w).

To estimate missing data, we developed the following procedure:

- For couples (i, w) having $coverage_i^{[w]} < 0.85$ as defined by Equation (4.9), we defined a linear model

$$\begin{aligned} boarded_i^{[w]} &= \beta_{1,i} * n\_total\_trains_i^{[w]} + \beta_{2,i} * n\_total\_tickets\_boarded_i^{[w]} \\ dropped_i^{[w]} &= \beta_{3,i} * n\_total\_trains_i^{[w]} + \beta_{4,i} * n\_total\_tickets\_dropped_i^{[w]} \end{aligned} \tag{4.10}$$

Where $n\_total\_tickets\_boarded_i^{[w]} = \sum_{j=1}^{J} t_{ij}^{*[w]}$, $n\_total\_tickets\_dropped_i^{[w]} = \sum_{i=1}^{I} t_{ij}^{*[w]}$ and $n\_total\_trains_i^{[w]}$ is the number of trains that stopped at station i during week w (i.e., the sum of train rides having APC states in $[-1, 0, 1, 2, 3, 4, 5, 6]$). The ticket-derived seed-matrices $T^{*[w]}$ are symmetrical, thus in our application $n\_total\_tickets\_boarded_i^{[w]} = n\_total\_tickets\_dropped_i^{[w]}$ $\forall i \in S$, $\forall w \in W$. However, to generalise our process, we keep the distinction between $n\_total\_tickets\_boarded_i^{[w]}$ and $n\_total\_tickets\_dropped_i^{[w]}$. Notice that the notation $n\_total\_tickets$ refers to the number of trips estimated in the tickets' conversion process at the previous step and not to the number of tickets and subscriptions purchased in station i and week w.

The training of the model is performed by applying the reweighting defined by

$$\begin{aligned} n\_total\_tickets\_boarded_i^{[w]} &= n\_total\_tickets\_boarded_i^{[w]} * coverage_i^{[w]} \\ n\_total\_tickets\_dropped_i^{[w]} &= n\_total\_tickets\_dropped_i^{[w]} * coverage_i^{[w]} \\ n\_total\_trains_i^{[w]} &= n\_valid\_trains_i^{[w]} \\ boarded_i^{[w]} &= partial\_boarded_i^{[w]} \\ dropped_i^{[w]} &= partial\_dropped_i^{[w]} \end{aligned}$$

We reweight the tickets and the number of trains using percentage $coverage_i^{[w]}$, and use actual (not reweighted) data when making predictions considering the total number of trains $n\_total\_trains_i^{[w]}$. This reweighting is needed since the outcome of training data is partial: $boarded\_partial_i^{[w]}$ and $dropped\_partial_i^{[w]}$ account only for boarded and dropped passengers of the trains having valid data, which are a percentage $coverage_i^{[w]}$ of total trains.

In the case of station Verona Porta Nuova, which has no ticket data, the model assumes the form

$$\begin{aligned} boarded_i^{[w]} &= \beta_{1,i} * n\_total\_trains_i^{[w]} \\ dropped_i^{[w]} &= \beta_{1,i} * n\_total\_trains_i^{[w]} \end{aligned} \tag{4.11}$$

Notice that we removed intercepts $\beta_{0,i}$ in the models since intercepts were often negative and caused negative predictions of $boarded_i^{[w]}$ and $dropped_i^{[w]}$ in a considerable number of cases. We observed that forcing intercepts at zero decreases the number of negative predictions.

- For couples (i, w) having $coverage_i^{[w]} \geq 0.85$, we rescaled counter data as

$$\begin{aligned} boarded_i^{[w]} &= \frac{partial\_boarded_i^{[w]}}{coverage_i^{[w]}} \\ dropped_i^{[w]} &= \frac{partial\_dropped_i^{[w]}}{coverage_i^{[w]}} \end{aligned} \tag{4.12}$$

  We decide to apply rescaling when $coverage_i^{[w]} \geq 0.85$ for two reasons:

  1. Data having high coverage need only minor corrections in the number of estimated passengers. The limit case is $coverage_i^{[w]} = 1$, corresponding to all trains having valid data and no need to estimate via model. We decide not to apply the prediction model when $coverage_i^{[w]} \geq \tau$, where τ indicates a rescaling threshold. In the case where $\tau \leq coverage_i^{[w]} < 1$, the estimates for trains having missing counter data come from the rescaling process defined in Equation (4.12).
  2. When we experimented with not applying rescaling, we noticed that for high values of $coverage_i^{[w]}$, the prediction model often gave estimates of $boarded_i^{[w]}$ and $dropped_i^{[w]}$ lower than the (partial) initial data $partial\_boarded_i^{[w]}$ and $partial\_dropped_i^{[w]}$. We thus decide to apply rescaling and tune threshold τ to minimise the problem described. The chosen threshold is $\tau = 0.85$.

- For couples (i, w) of inactive stations having $n\_total\_trains_i^{[w]} = 0$, we set

$$\begin{aligned} boarded_i^{[w]} &= 0 \\ dropped_i^{[w]} &= 0 \end{aligned}$$

  Since no trains have stopped at station i and week w, no passengers can board or drop at that station, thus we force predictions to 0.

In the end, we round the final results since $boarded_i^{[w]}$ and $dropped_i^{[w]}$ have a natural interpretation as the number of boarded and dropped passengers for each week and station and should therefore be integers.

Aggregation of the Integrated Subscriptions area

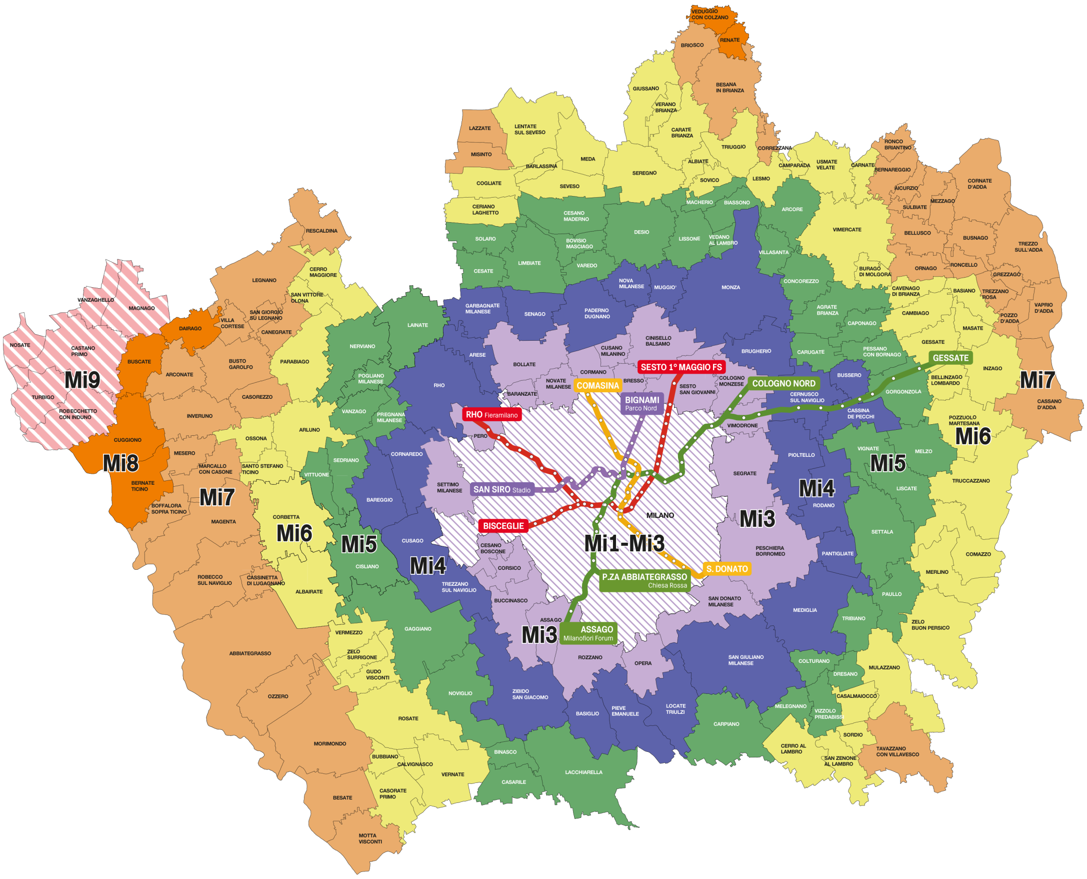


Figure 4.3: Integrated Subscriptions area

After producing the two inputs needed to perform the Furness method (seed OD matrices $T^{*[w]}$ from ticket data and margin vectors $\mathit{boarded}_i^{[w]}$ and $\mathit{dropped}_i^{[w]}$ from counter data), we aggregate the Integrated Subscriptions area (IS area, see Figure 4.3), covering Milan and Monza provinces. This choice is taken for two reasons:

- As can be seen from Figure 3.4, our study covers only a tiny fraction of the train lines in this area. Every estimation of train movements in the area from our data will be heavily underestimated, and this would lead to significant problems in the computation of mobility-based spatial weights, where it is crucial to estimate correctly the proportion of trips happening between every area of the study.
- Every OD matrix estimation will suffer from the missing ticket data internal to Milan. This problem could be overcome in the application, as we will explain in the next step of the pipeline. Still, another advantage of aggregating the IS area is to resolve this matter without applying the correction of zero cells explained in subsection 4.1.1.

We thus deem our data insufficient to estimate movements internal to the IS area. We can only assess trends between this area and stations external to it.

For this reason, our analysis from now on is conducted considering the IS area aggregated into one point (identified by the name "IS area" in the analysis and graphs from now on). Thus, the number of zones in our portion of the Trenord railway network where trips can begin or end reduces from 46 (the initial stations) to 32 (31 stations and the aggregated IS area).

Therefore, for every station belonging to the IS area (i.e., stations Arcore, Carnate Usmate, Cassano d'Adda, Melzo, Milano Centrale, Milano Greco Pirelli, Milano Lambrate, Milano Porta Garibaldi, Milano Villapizzone, Monza, Pioltello Limito, Pozzuolo Martesana, Sesto S. Giovanni, Trecella and Vignate), we compute

$$\begin{aligned} boarded_{IS}^{[w]} &= \sum_{i \in IS} boarded_i^{[w]} \\ dropped_{IS}^{[w]} &= \sum_{i \in IS} dropped_i^{[w]} \\ t_{i,IS}^{*[w]} &= \sum_{j \in IS} t_{ij}^{*[w]} \\ t_{IS,j}^{*[w]} &= \sum_{i \in IS} t_{ij}^{*[w]} \end{aligned}$$

Movements internal to the IS area are allowed in the IS aggregated case. Thus, the final OD matrices will have $t_{IS,IS}^{[w]} \neq 0$.

The aggregation of tickets generates entries $t_{IS,IS}^{[w]} \neq 0$. These entries are partial estimates and only represent a portion of the trips that occur within the IS area. The entries are based on trips where both the origin and destination are within the IS area, with one of the stations being located outside Milan. It is important to note that $t_{IS,IS}^{[w]}$ does not include trips that both start and end within Milan. Despite its limitations, we will use this entry as a starting point for the Furness method, since we are unable to provide a precise estimate of the number of trips that originate and end within the IS area using ticket data alone.

After this operation, we can estimate OD matrices thoroughly describing the railway movements in the study area, except for movements between the IS area and station Treviglio because they are connected by train lines RE2, RE6, R4, and suburban lines S5 and S6 (see Figure 3.4) about which we have no data. The estimated trips in this path will thus suffer from some underestimation.

Application of Furness method

To obtain the weekly OD matrices describing the number of trips $t_{ij}^{[w]}$ happening during week w from station i to station j, we apply the Furness algorithm described in subsection 4.1.1 separately to every one of the 37 weeks of the study. The ticket OD matrices $T^{*[w]}$ are the seed matrices for the algorithm, while the margins are the vectors of boarded passengers in station i $boarded_i^{[w]}$ as $p_i^{[w]}$ and the number of passengers dropped at station j during week w $dropped_j^{[w]}$ as $a_j^{[w]}$, considering the IS area aggregated into a single zone.

Before applying the Furness algorithm, we check the concordance of each of the 37 seed matrices $t_{ij}^{*[w]}$ with margin vectors $p_i^{[w]}$ and $a_j^{[w]}$ through the Wald test explained in subsection 4.1.1. The test's null hypothesis is H_0 : *Seed matrix* T^* *agrees with margins* $\{p_i, a_j\}$. Suppose we do not reject the null hypothesis. In that case, the assumptions we defined in the ticket conversion procedure successfully obtained seed OD matrices coherent with the number of total boarded and dropped passengers for each station and week.

Because of the estimation process and errors in the APC system, every week $w \in W$ has $\sum_{i=1}^{I} p_i^{[w]} \neq \sum_{j=1}^{J} a_j^{[w]}$. This contradicts the assumption expressed by Equation (4.1). We thus rescale tickets, boarded and dropped passengers, adopting the probability interpretation defined by Equation (4.8), as

$$\begin{aligned} \pi_{ij}^{*[w]} &= \frac{t_{ij}^{*[w]}}{\sum_{(i,j)\in S} t_{ij}^{*[w]}} \\ \rho_i^{[w]} &= \frac{boarded_i^{[w]}}{\sum_{k\in S} boarded_k^{[w]}} \\ \alpha_j^{[w]} &= \frac{dropped_j^{[w]}}{\sum_{k\in S} dropped_k^{[w]}} \end{aligned} \tag{4.13}$$

After this operation $\sum_{i\in S} \rho_i^{[w]} = \sum_{j\in S} \alpha_j^{[w]} = 1 \ \forall w \in W$ and $\sum_{(i,j)\in S} \pi_{ij}^{*[w]} = 1 \ \forall w \in W$. Then, Furness method is applied and the resulting matrix $\pi_{ij}^{[w]}$ has to be corrected as follows

$$t_{ij}^{[w]} = \pi_{ij}^{[w]} \sum_{k\in S} boarded_k^{[w]}$$

to recover the final matrix $T^{[w]}$. We could alternatively recover the final matrix by multiplying by the total number of dropped passengers in week w.

We also have to deal with the zero correction problem to apply the Furness method. Since the ticket number is vastly underestimated compared to the number of actual passengers travelling from and to each station, some couples of stations (i, j) might have a 0 in the ticket OD matrices $t_{ij}^{*[w]}$, but some movements have happened between them. Furness cannot modify 0 entries, so we apply the procedure described in subsection 4.1.1 to correct 0 values. This procedure develops as follows:

1. We substitute every zero entry of the seed matrix $t_{ij}^{*[w]}$ with a 1, except for entries $t_{ii}^{*[w]}$ corresponding to couples of the same station. An exception is the IS area for which $t_{IS,IS}^{*[w]} \neq 0$. This correction is applied before the rescaling described in Equation (4.13).
2. We apply Furness algorithm and round the resulting matrix $T^{[w]}$.
3. We perform the binomial test described in Equation (4.7) for every cell which was initially zero. The number of tests performed equals the number of zero cells (excluding the diagonal) of matrix $t_{ij}^{*[w]}$.
4. The p-values of all the binomial tests are corrected by applying the False Discovery Rate controlling procedure.
5. We put ticket cells t_{ij}^{*} for which we cannot reject the null hypothesis at level $\alpha = 0.05$ to zero. These are the couples (i, j) which we suppose have no movements from i to j during week w.
6. We perform Furness method again to adjust the corrected seed matrix to the margin vectors.

We have some inactive stations i with no trains in week w ($n\_total\_trains_i^{[w]} = 0$), thus $p_i^{[w]} = 0$ and $a_i^{[w]} = 0$. This does not cause problems in the application of Equation (4.5): the equation will not be applied in such cases and these inactive stations will correspond to an empty row and column in the final matrix $T^{*[w]}$.

At the end of the procedure, we round the final matrix $T^{[w]}$ since its entries $t_{ij}^{[w]}$ have a natural interpretation as the number of trips starting from station i and ending in station j during week w.

4.2. Exploratory analysis of Trenord data

We first present an exploratory analysis conducted on the two datasets provided by Trenord, described in subsection 3.1.2 to highlight some criticalities in the data. Figure 4.4 displays the number of tickets purchased during the study period and the additional months of February and August, distinguishing ticket types.

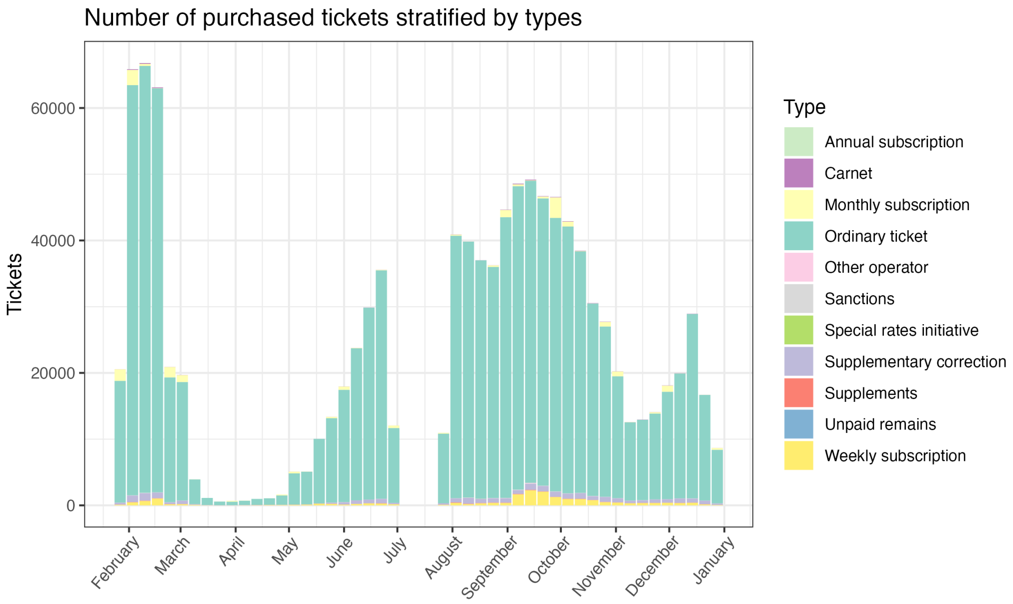


Figure 4.4: Number of tickets purchased during the study period, stratified by types

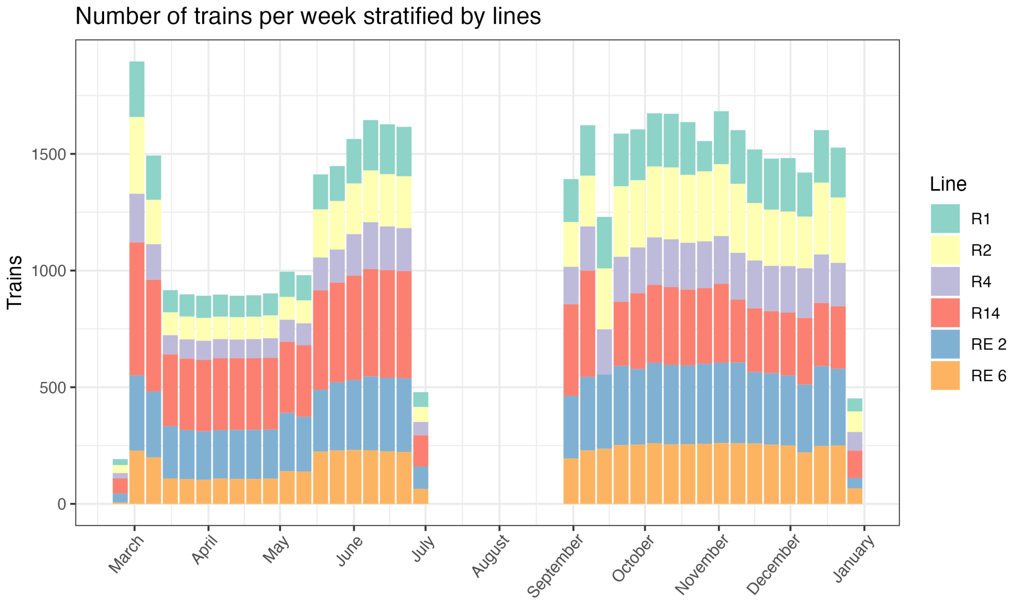


Figure 4.5: Number of trains per week, stratified by lines

We can notice that ordinary tickets are the most purchased kind of ticket, followed by monthly and weekly subscriptions. Moreover, the purchases greatly decrease during the lockdown period in March and April and then increase after the easing of some restrictions on May 4. During the red zone or second lockdown period in November 2020, purchases decrease again.

Considering counter data, Figure 4.5 displays the number of trains analysed in our study. The number of train rides reduces during the first lockdown period because of restrictions to public transport. Indeed, Trenord cut the scheduled train rides by 40% [26], while the second lockdown period saw a reduction of capacity by 50% [33], but little decrease in the number of train rides.

We can notice that there are weeks when the number of trains is much smaller than other weeks because we do not have full data describing all the 7 days of the week. These are weeks number 08 (which is composed of just a day, March 1, since the previous days of the week belong to February and have no counter data), 26 (composed of just two days, June 29 and 30, since then June ends and July does not belong to our study period), 35 (composed by six days, September 1, 2, 3, 4, 5, and 6), and 52 (formed by 4 days, December 28, 29, 39 and 31). Week 37 in September also has no trains for the R14 line.

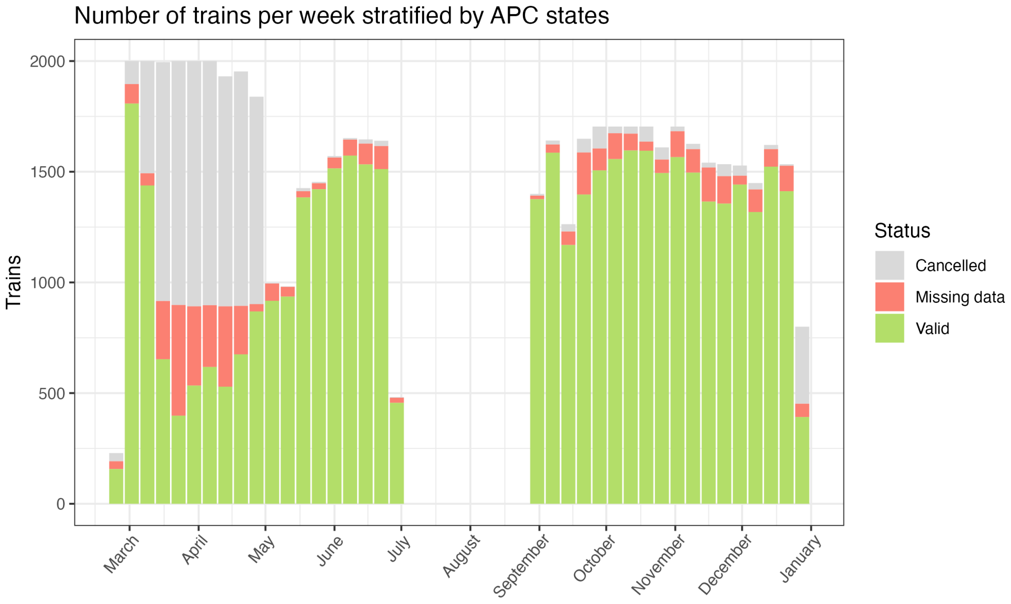


Figure 4.6: Number of trains per week, stratified by counter states

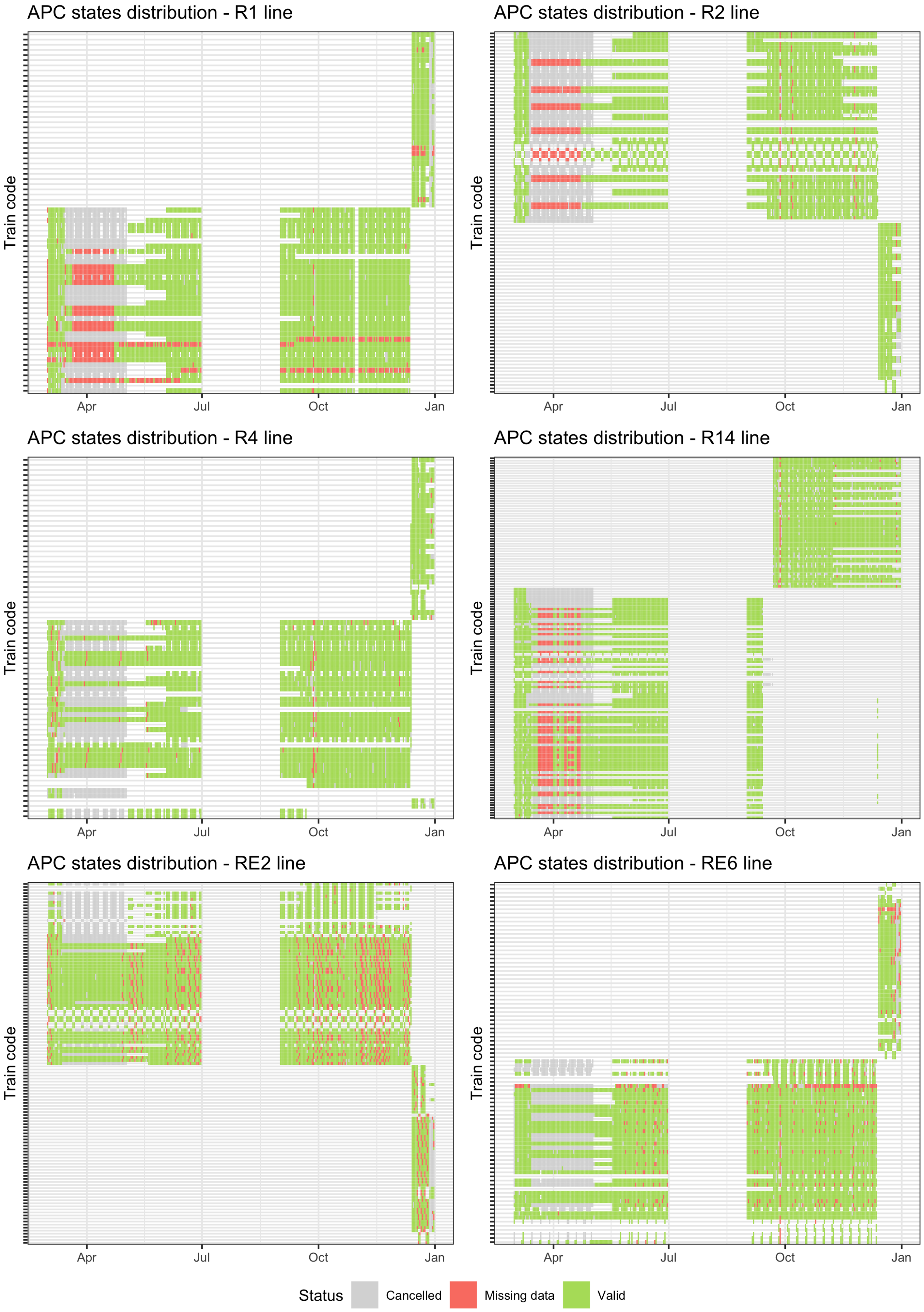


Figure 4.7: APC states distribution in the six train lines

Figure 4.6 displays the number of trains per week divided by counter states, grouped into cancelled (APC states in $[-5, -4, -3, -2]$), missing (states $[-1, 5, 6]$) or valid (states $[0, 1, 2, 3, 4]$). Most of our dataset consists of valid data, i.e., trains with a successful counting process through equipped or partially equipped APC train rides or interpolations. The first lockdown period shows more cancelled trains and trains having missing counter data than the rest of the year, which we should deal with in our application.

Table 4.2 summarises the distribution of APC states in the dataset. States -4, -3, and 5 are not present in the dataset and are thus not considered in the analysis.

APC states	Number of train rides
-5	38
-4	0
-3	0
-2	535
-1	169
0	560
1	10351
2	662
3	87
4	365
5	0
6	736

Table 4.2: APC states of the train rides of the study

Data concerns six train lines: R1, R2, R4, R14, RE2, and RE6. Figure 4.7 shows the distribution of valid, missing, and cancelled data for each train line.

Lastly, Table 4.3 summarises the percentage of missing data between the six train lines.

Train rides with missing data are much more likely to happen during the lockdown period. Moreover, lines R1, R2 and R14 have several weeks where no valid data has been registered.

Lines	Percentage of missing train rides
R1	12.8%
R2	8.7%
R4	2.3%
R14	8.9%
RE2	12.6%
RE6	5.8%

Table 4.3: Percentage of missing counter data by line, excluding cancelled trains

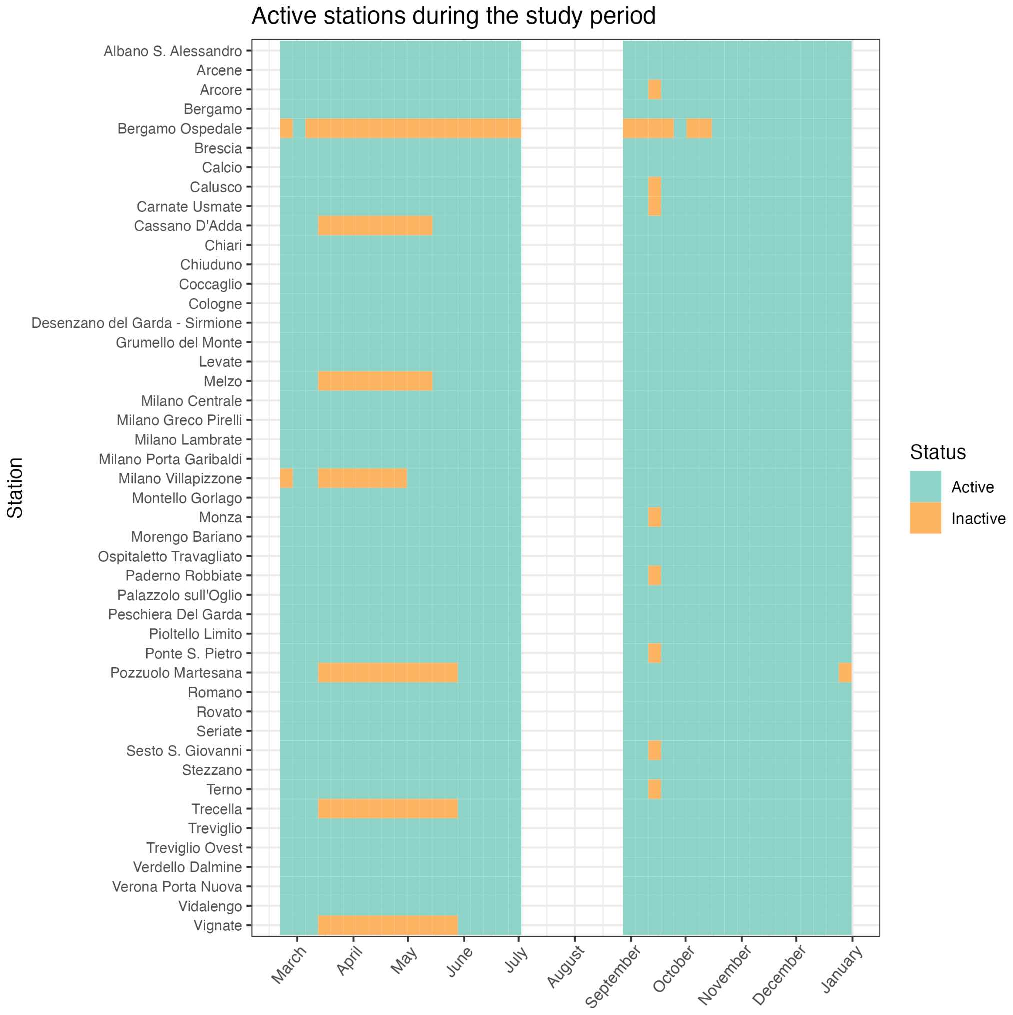
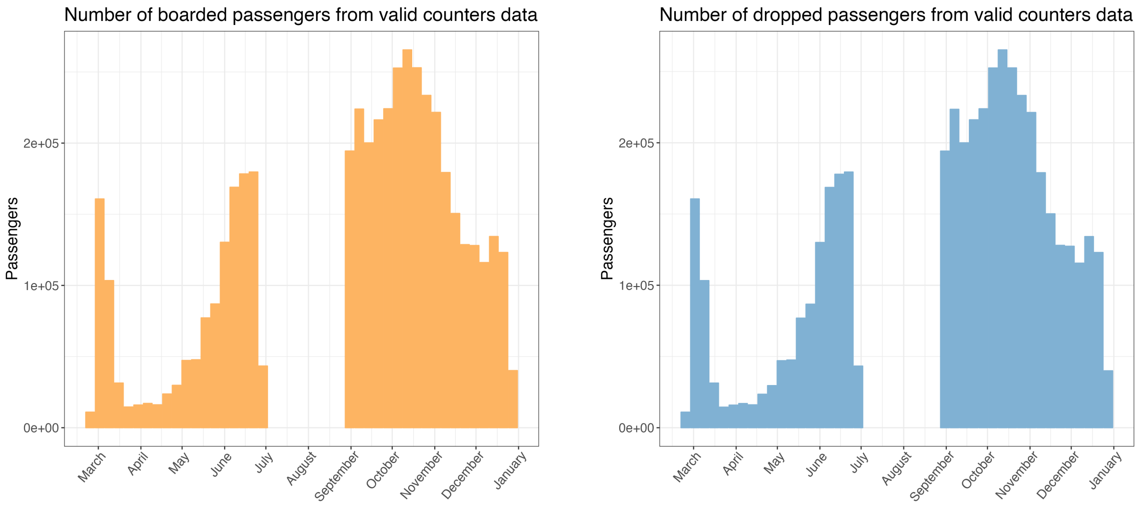


Figure 4.8: Stations' activity states in 2020

Considering stations, some of them had no train stopping there during certain weeks.

Figure 4.8 shows the inactive stations. Station Bergamo Ospedale has been inactive for most of 2020, and stations Cassano d'Adda, Melzo, Milano Villapizzone, Pozzuolo Martesana, Trecella and Vignate also experience several weeks of inactivity.

Figure 4.9 displays the number of boarded and dropped passengers considering only valid counters data.



(a) Boarded passengers

(b) Dropped passengers

Figure 4.9: Total number of valid boarded and dropped passengers per week

We notice that the number of boarded and dropped passengers decreases during the first lockdown, then increases after the easing of the restrictions and slightly reduces during the second lockdown. When looking at the graphs, remember that the first weeks of March and September and the last of June and December are partial (they are not composed of seven days of data) and should thus be interpreted carefully.

Finally, we notice that the number of partial boarded passengers summed for all stations (which should theoretically be equal for every week) never equals the number of partial boarded passengers summed for all stations. There is always a difference, with the number of partial boarded passengers always greater than the number of partial dropped passengers. Figure 4.10 shows the difference between the two quantities, which ranges from 22 to 686. This problem needs to be handled during the application of the Furness method, since it contradicts the assumption in Equation (4.1).

We now recap some key points shown in the two datasets in this exploratory analysis:

- We have evidence of a strong decrease in purchased tickets, number of train rides and number of passengers boarded and dropped during the first and (to a lesser

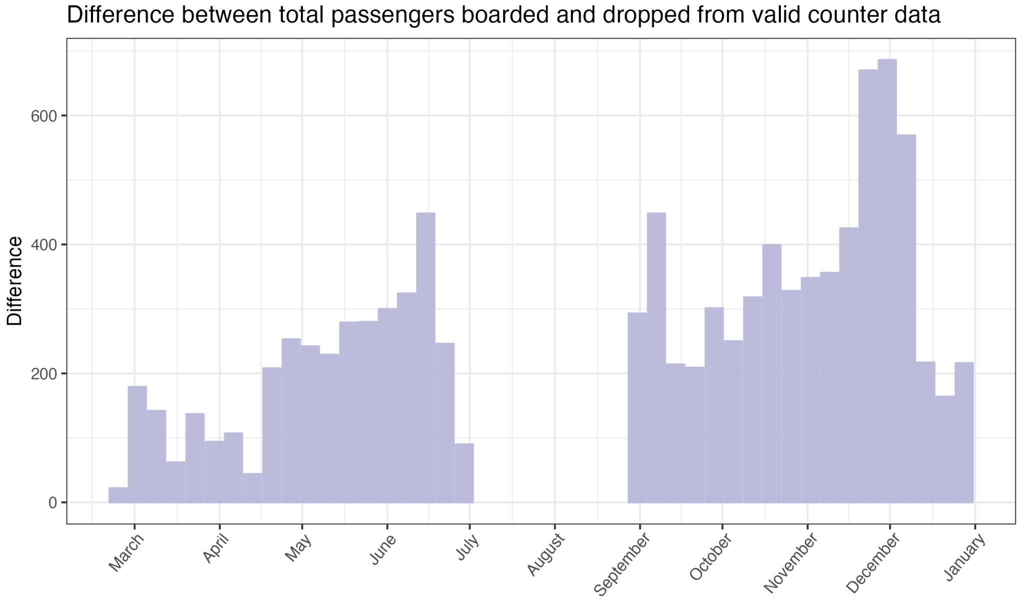


Figure 4.10: Difference between total passengers boarded and dropped from valid counter data

extent) second lockdown period. Thus, we expect similar reductions of movements in the estimated OD matrices in the same periods.

- Most of the counter dataset consists of valid data, but we have some missing data distributed between the six train lines. Train rides with missing counter data are more frequent during the first lockdown and a critical step of the pipeline is to correct the partial passenger counts accounting for these missing data.
- The inactive stations of Figure 4.8 will correspond to no movements ending and starting from these couples of station and week in the estimated OD matrices. This point is underlined because of its consequences on the definition of the mobility-based spatial description in the spatial analysis of Chapter 5.

4.3. Results

In this Section, we apply the methodology described in Section 4.1 to obtain 37 weekly OD matrices describing movements through a limited portion of the Trenord railway network covering six train lines in 8 months of 2020.

We recap briefly the procedure to obtain the weekly estimated OD matrices, and then show results of its application for each step:

1. We convert ticket data into ticket-estimated OD matrices $T^{*[w]}$, using various as-

sumptions to convert each ticket type into one or more trips between couples of stations (i, j) in a week w. We apply this process to each of the 46 stations S and each of the 37 weeks W in the study.

2. We derive margin vectors that describe the total number of boarded $(p_i^{[w]})$ and dropped $(a_i^{[w]})$ passengers for each station $i \in S$ and each week $w \in W$. To obtain these quantities, we aggregate the counter data for each train ride equipped with APC systems or interpolations and then define a model to correct passenger counts accounting for trains having missing counter data.
3. We aggregate the 15 stations of the IS area.
4. We perform the Furness algorithm, which takes the seed ticket matrix $T^{*[w]}$ and the margin vectors $p_i^{[w]}$ and $a_i^{[w]}$ as input to derive the final estimates $T^{[w]}$ for every $w \in W$. Each cell $t_{ij}^{[w]}$ of these matrices represents the number of trips by train from station i to station j during week w.

4.3.1. Conversion of ticket data into seed OD matrices

Applying the assumptions described in subsection 4.1.2, we obtain 37 ticket-estimated OD matrices $T^{*[w]}$ describing movements estimated by tickets and subscriptions purchased in 2020. These matrices will be the input of the Furness method, which will correct them by multiplying each cell for suitable values to make their rows and columns sum to the number of boarded and dropped passengers.

Figure 4.11 reports an example of ticket matrix $(T^{*[09]})$ estimated by applying the assumptions described. Notice the absence of any ticket-estimated trip to or from Verona Porta Nuova and of tickets internal to Milan, which has stations Milano Lambrate, Milano Centrale, Milano Porta Garibaldi, Milano Villapizzone, Milano Greco Pirelli. The emptiness of these cells is caused by the missing ticket data for these areas.

We detect many movements around stations Brescia, Treviglio, and Bergamo in each of the 37 matrices, which are some of the busiest station in the network. Moreover, we observe a strong decrease in all the cells during the first lockdown period and another reduction during the second one.

4.3.2. Counter data aggregation and estimation of missing data

Figure 4.12 graphically recaps the procedure to estimate the margin vectors $boarded_i^{[w]}$ and $dropped_i^{[w]}$ from partial counter data received from train rides with valid passengers

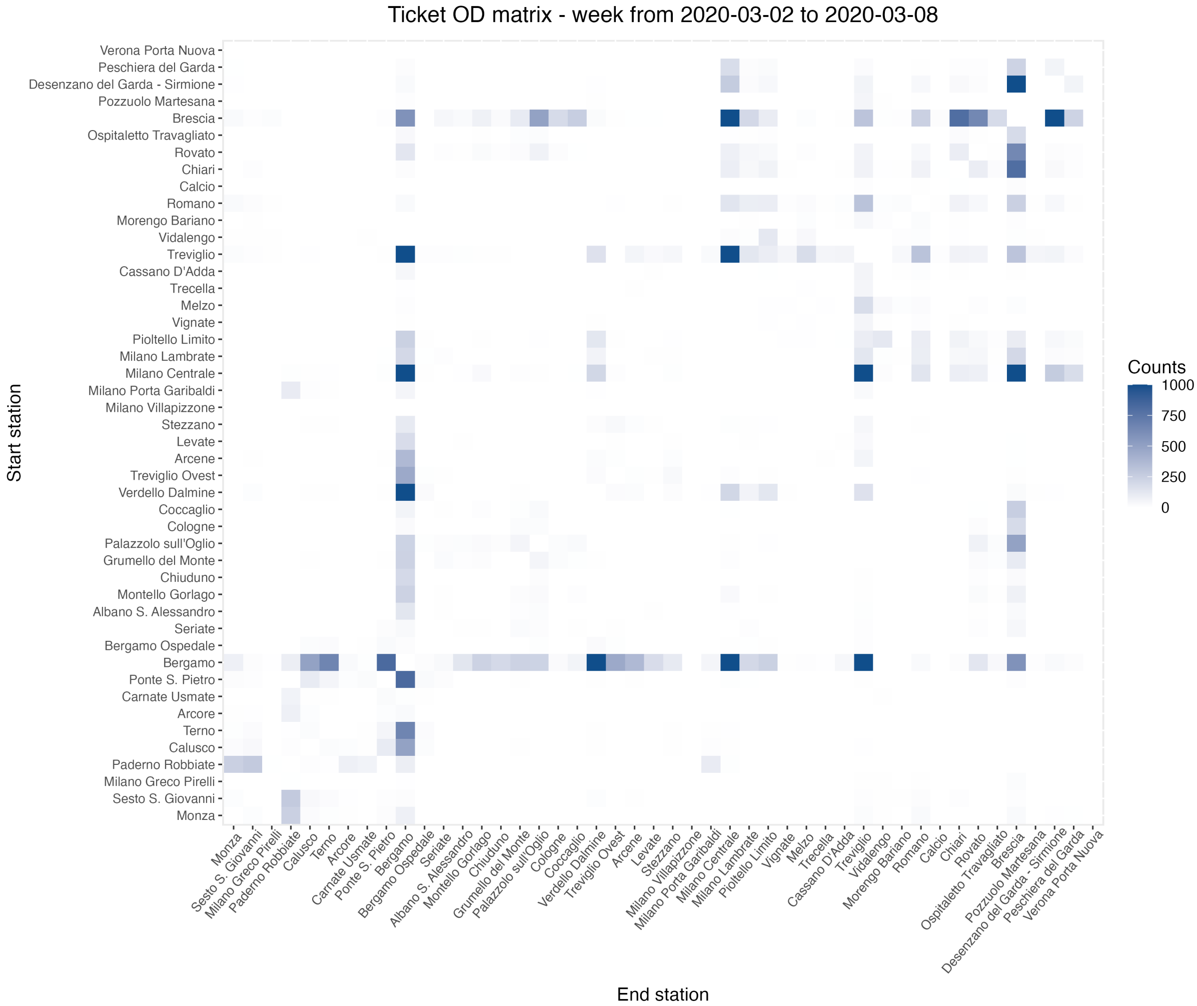


Figure 4.11: Ticket matrix $T^{*[09]}$

counts.

First, we comment the regression model expressed in Equation (4.10) and Equation (4.11). The models achieve an R^2 which varies a lot between stations, as shown in Figure 4.13.

R^2 of the models (considering the reweighted training data) is 0.83 for the boarded passengers model and 0.85 for the dropped passengers one. Thus, variables $n\_total\_trains_i^{[w]}$, $n\_total\_tickets\_boarded_i^{[w]}$ and $n\_total\_tickets\_dropped_i^{[w]}$ are significant in explaining the outcomes. Lines going from Brescia to Verona, from Bergamo to Brescia and from Treviglio to Bergamo show high values of R^2. At the same time, the coefficient lowers in the area outside Bergamo covered by the R14 line as well as in the area from Treviglio to Milan, where we can explain this behaviour in the fact that there exist other suburban lines (S5 and S6) travelling from Treviglio to Milan for which passengers may have bought tickets. We notice the strange case of stations Vidalengo and Chiari, which have a high

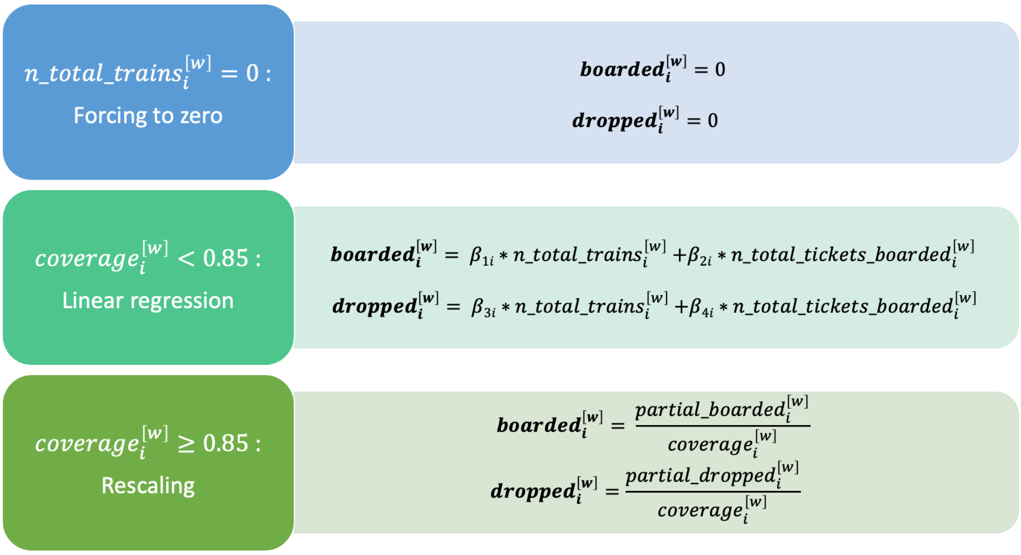

$n\_total\_trains_i^{[w]} = 0$:
Forcing to zero
$boarded_i^{[w]} = 0$
$dropped_i^{[w]} = 0$
$coverage_i^{[w]} < 0.85$:
Linear regression
$boarded_i^{[w]} = \beta_{1i} * n\_total\_trains_i^{[w]} + \beta_{2i} * n\_total\_tickets\_boarded_i^{[w]}$
$dropped_i^{[w]} = \beta_{3i} * n\_total\_trains_i^{[w]} + \beta_{4i} * n\_total\_tickets\_boarded_i^{[w]}$
$coverage_i^{[w]} \geq 0.85$:
Rescaling
$boarded_i^{[w]} = \frac{partial\_boarded_i^{[w]}}{coverage_i^{[w]}}$
$dropped_i^{[w]} = \frac{partial\_dropped_i^{[w]}}{coverage_i^{[w]}}$

Figure 4.12: Estimation of boarded and dropped passengers for each week $w \in W$ and station $i \in S$ from partial counter data

R^2 for the model predicting boarded passengers but a much lower one for the dropped passengers model.

Notice that tickets internal to the Milan municipality are not reported in our data. Thus, stations in this area suffer from an underestimation in the number of trips estimated by tickets. Because of this, estimates $\beta_{2,i}$ and $\beta_{4,i}$ are usually higher for these stations. The same happens for stations on the RE6 line, like Peschiera del Garda or Desenzano del Garda, to account for the missing tickets to and from Verona Porta Nuova.

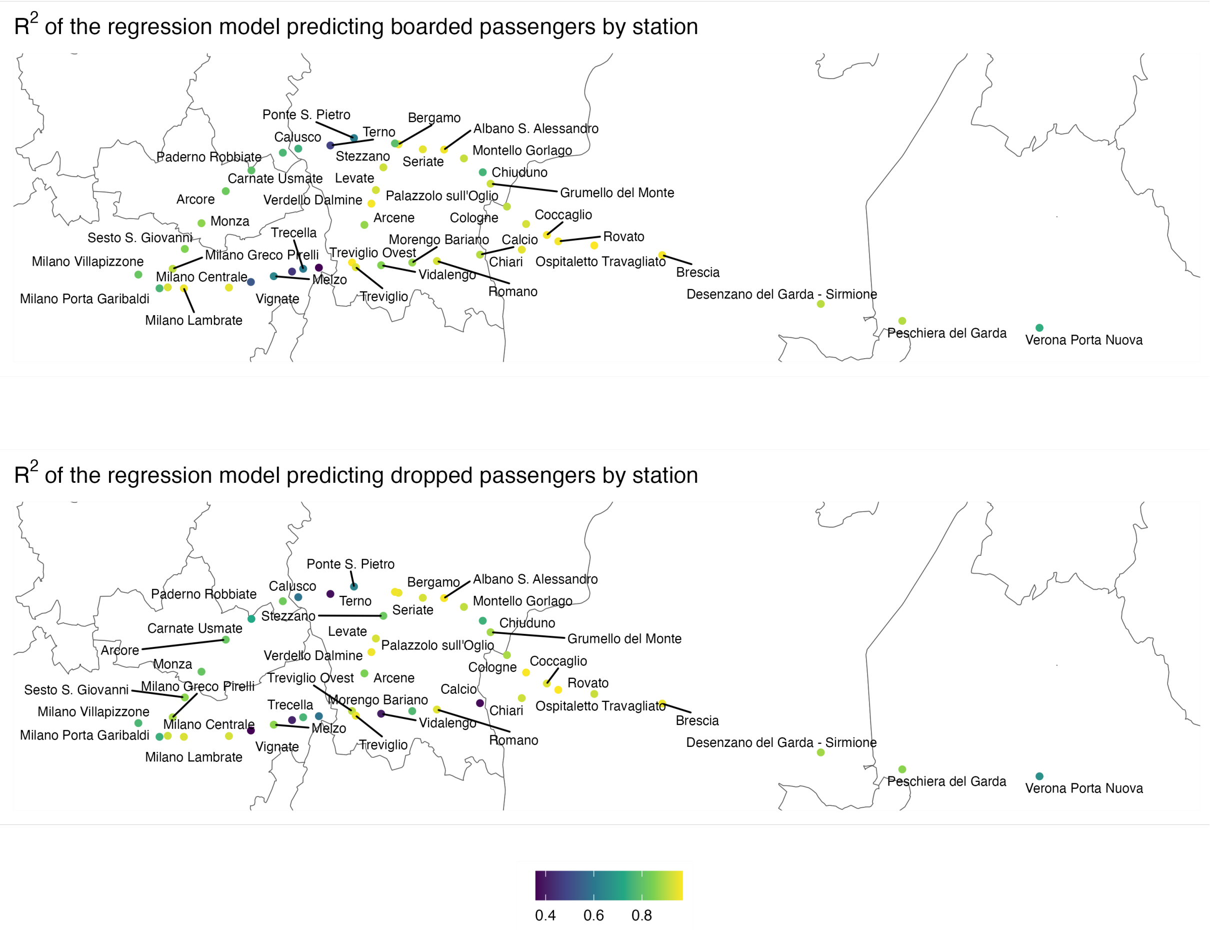
We apply the model of Figure 4.13 to every station $i \in S$ to estimate $boarded_i^{[w]}$ and $dropped_i^{[w]}$ $w \in W$. Figure 4.14 shows an example of the result of the margins estimation process in station Bergamo.

Applying the model choices described in subsection 4.1.2, the final prediction model shows some desirable properties:

1. The model never predicts negative values for $boarded_i^{[w]}$ or $dropped_i^{[w]}$.

2. The model is found to have the minimum number of underpredictions, i.e., couples (i, w) where $boarded_i^{[w]} < partial\_boarded_i^{[w]}$ or $dropped_i^{[w]} < partial\_dropped_i^{[w]}$. The tuning of the threshold τ to apply rescaling or linear regression to the final value of $\tau = 0.85$ has been critical in reducing this phenomenon.

3. The model can fill data for the lockdown period, where some stations have



R^2 of the regression model predicting boarded passengers by station

R^2 of the regression model predicting dropped passengers by station

Figure 4.13: R^2 of the linear regression models estimating boarded and dropped passengers by station

$coverage_i^{[w]} = 0$ for several weeks. The predictions have a similar trend to other stations having valid data for which rescaling is applied.

However, some issues in the missing data estimation process remain:

1. There still are some stations showing underpredictions ($boarded_i^{[w]} < partial\_boarded_i^{[w]}$ or $dropped_i^{[w]} < partial\_dropped_i^{[w]}$). These stations are Vignate, Cassano d'Adda, Melzo, Milano Centrale and Trecella. Figure 4.15 shows an example concerning station Vignate. Luckily, this problem usually happens for stations moving only a small number of passengers, except for Milano Centrale. The problem occurs more often in stations belonging to the Integrated Subscriptions area covering Milan and Monza provinces, which are then aggregated in the next step of the analysis (see subsection 4.1.2). We thus resolve to keep these underpredictions since they would not decisively influence the next steps of the pipeline.

2. The model sometimes predicts a peak at the beginning of lockdown in cases where $coverage_i^{[w]} < 0.85$ and the regression model plugs in. These predictions are likely caused by the number of trips estimated by tickets for these weeks, which remains similar to those preceding the lockdown. It is plausible that these trips estimated from tickets come from subscriptions bought before the beginning of the lockdown, which have been purchased but seldom used. Figure 4.16 shows an example of this problem in station Brescia. For stations with $coverage_i^{[w]} > 0.85$ in the period, where the model does not plug in, we see that real data never show peaks and the number of passengers radically drops at the beginning of the lockdown. We thus suspect that the model provides inaccurate predictions, but we did not find a way to solve this issue. 9 out of 46 stations show this behaviour: Arcore, Brescia, Calusco, Carnate Usmate, Cassano d'Adda, Chiuduno, Montello Gorlago, Paderno Robbiate and Rovato.

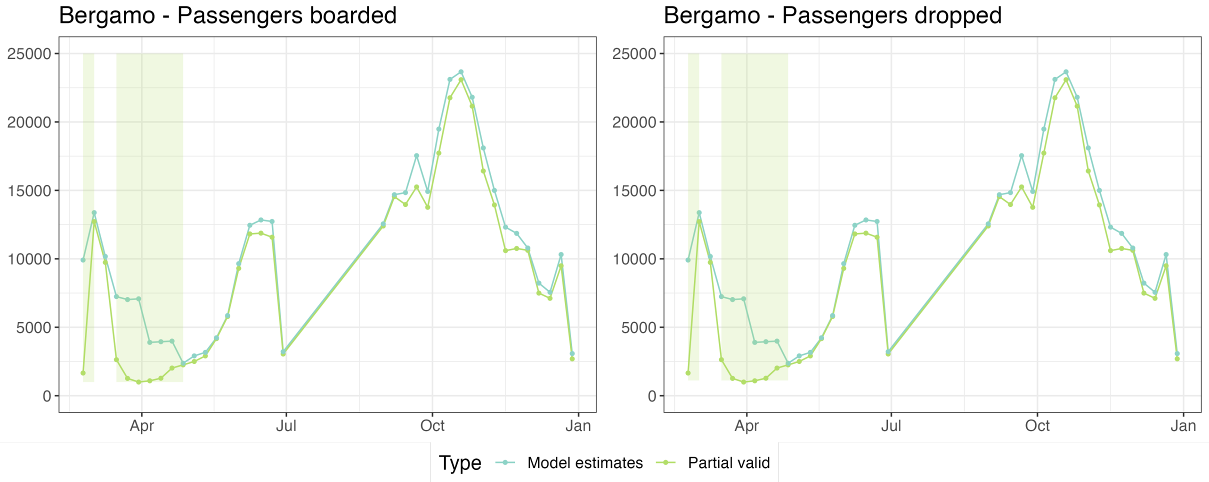


Figure 4.14: Estimates of boarded and dropped passengers in Bergamo station. The graph compares the model estimates (dark green) with the partial number of boarded and dropped passengers estimated by valid counter data (light green). Light green areas represent weeks where $coverage_i^{[w]} < 0.85$ and the regression model plugs in.

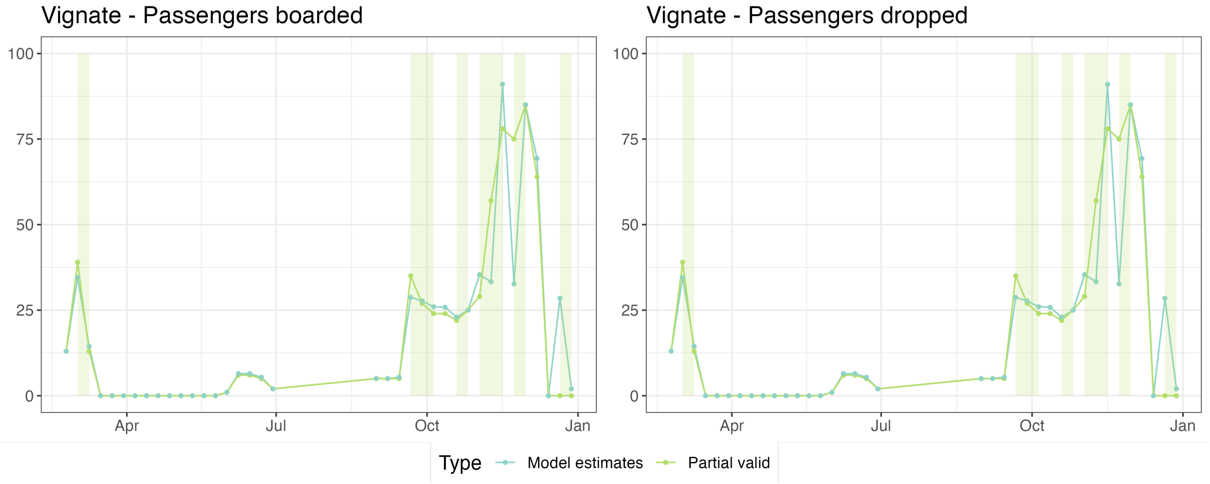


Figure 4.15: Estimates of boarded and dropped passengers in Vignate station. The graph compares the model estimates (dark green) with the partial number of boarded and dropped passengers estimated by valid counter data (light green). Light green areas represent weeks where $coverage_i^{[w]} < 0.85$ and the regression model plugs in.

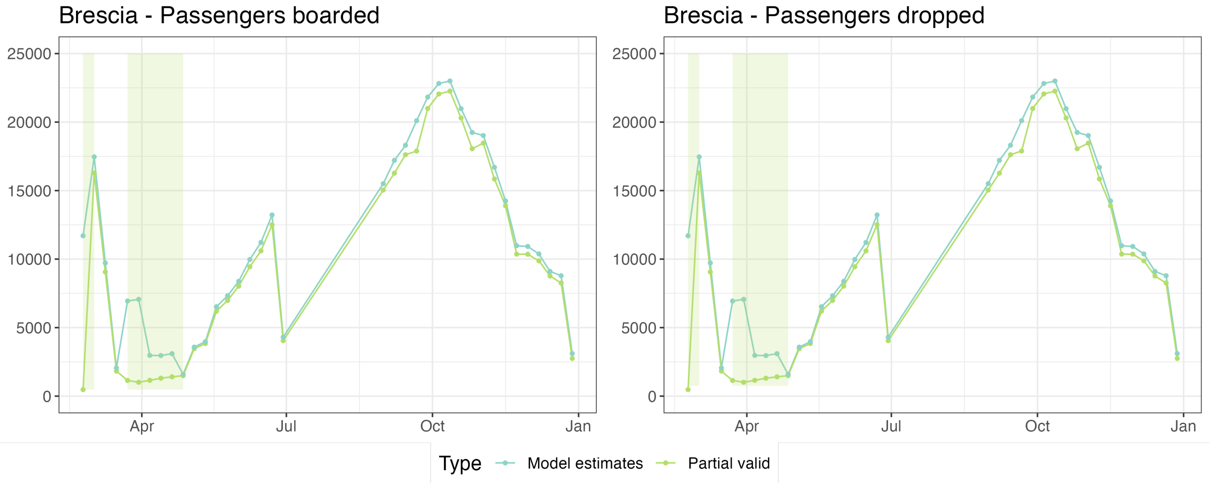


Figure 4.16: Estimates of boarded and dropped passengers in Brescia station. The graph compares the model estimates (dark green) with the partial number of boarded and dropped passengers estimated by valid counter data (light green). Light green areas represent weeks where $coverage_i^{[w]} < 0.85$ and the regression model plugs in.

4.3.3. Aggregation of the IS area and application of Furness method

After the conversion of tickets into seed OD matrices $T^{*[w]}$ and the estimation of margin vectors describing the total number of boarded and dropped passengers $boarded_i^{[w]}$ as $p_i^{[w]}$

and $dropped_i^{[w]}$ as $a_i^{[w]}$ for each station $i \in S$ and week $w \in W$, we aggregate the 15 stations of the IS area as described in subsection 4.1.2.

After the aggregation, we can proceed to derive the estimated OD matrices $T^{[w]}$ describing movements by train starting from station i and ending in station j during week w.

Before applying the Furness algorithm, we check the concordance of each of the 37 seed matrices $t_{ij}^{*[w]}$ with margin vectors $p_i^{[w]}$ and $a_j^{[w]}$ through the Wald test explained in subsection 4.1.1. The test's null hypothesis is H_0 : *Seed matrix* T^* *agrees with margins* $\{p_i, a_j\}$. We find no case where the null hypothesis could be rejected at any reasonable significance level. We conclude that the seed matrices generated from tickets agree with the margins vector derived by the model estimation from counter data. The test gives us evidence of the validity of ticket translation into trips and the estimation of total boarded and dropped passengers procedures. We can then proceed to merge the information through the Furness method and obtain the estimated weekly OD matrices.

We report two examples of the final OD matrices: Figure 4.17 shows matrix $T^{[14]}$, while Figure 4.18 shows matrix $T^{[38]}$. We can notice how $T^{[14]}$ (which describes a week during the first lockdown period) contains many fewer movements than $T^{[38]}$, which represents a week at the end of September, a period with no mobility restrictions to limit COVID-19 spread.

First, we can notice that the final OD matrices are coherent with some reality-induced principles:

1. We can see many movements around the major centres of Bergamo, Brescia, and IS area.
2. Most movements for each station are directed to and from the IS area, with some exceptions. This makes sense since the aggregated IS area has a number of total movements which is more than four times greater than the second-most busy station, which is Verona Porta Nuova.
3. Movements increase for stations belonging to the same train line.
4. Stations having two or three lines connecting them are even more strongly connected.
5. Mobility decreases during lockdowns.
6. Mobility resumes around mid-May after the end of the first lockdown.
7. Mobility increases after summer (around mid-September).
8. Mobility decreases again during the second lockdown and Christmas period.

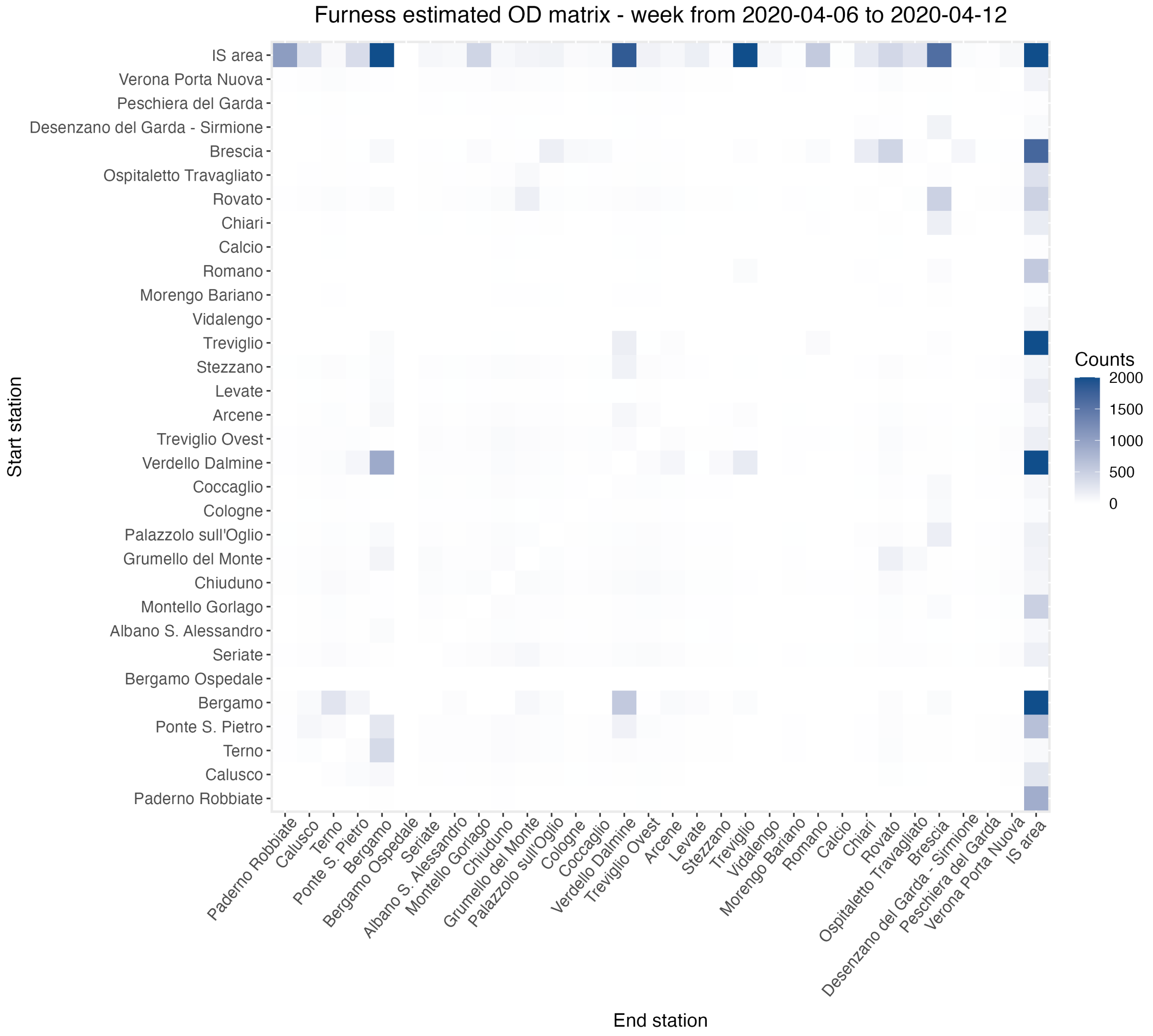


Figure 4.17: Furness estimated OD matrix $T^{[14]}$

Moreover, we can measure the concordance between the final matrix $t_{ij}^{[w]}$ and the margins $p_i^{[w]}$ and $a_j^{[w]}$, computing the margins errors ϵ_{row} and ϵ_{col} as the maximum deviation between each generated and desired margins as in Equation (4.6). Notice that since the final matrix is computed by multiplying the probability matrix $\pi_{ij}^{[w]}$ for the total number of boarded passengers, the row errors $\epsilon_{row}^{[w]}$ are much lower than the column errors $\epsilon_{col}^{[w]}$. Thus, we should consider the row margin errors when judging the goodness of fit of the procedure.

The row margin errors $\epsilon_{row}^{[w]}$ are consistently low, as shown in Figure 4.19. This is evidence that the margins of the estimated cells fit the actual margins $p_i^{[w]}$. Additionally, we observe that the margin errors $\epsilon_{row}^{[w]}$ are lower than when we experimented with not applying zero corrections, which is evidence in favour of applying the procedure.

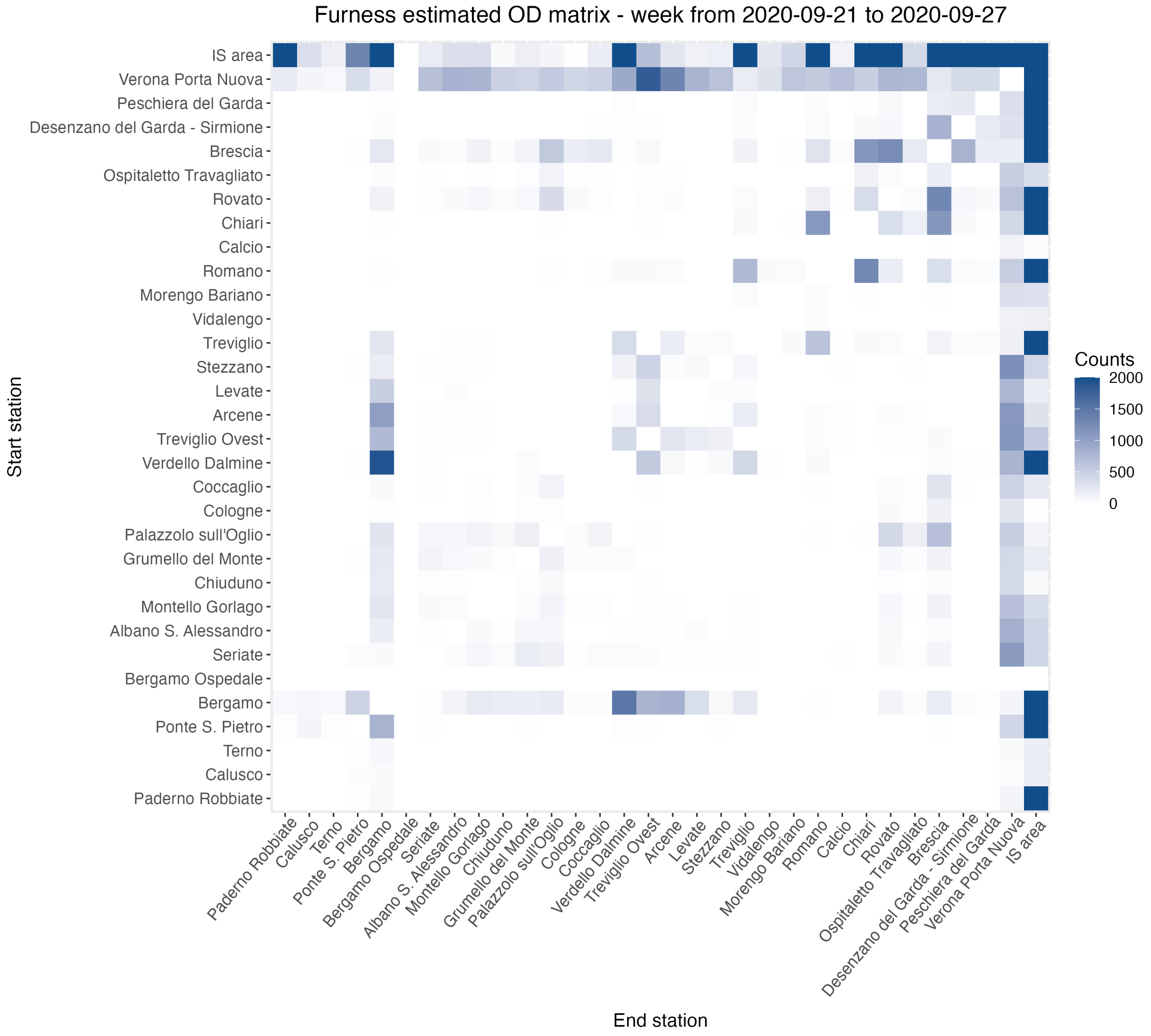


Figure 4.18: Furness estimated OD matrix $T^{[38]}$

One of the main critical issues observed in the final matrices concern station Verona Porta Nuova. This station is anomalous compared to all the others since we have no ticket data to and from it. Thus, all the entries in the final OD matrices have to be estimated by the zero correction procedure. Verona distributes its trips almost uniformly between all stations (as seen in Figure 4.18). Moreover, some minor stations have many movements toward Verona, even when Verona is quite far and does not belong to lines connecting these stations to Verona. An example is station Cologne; see Figure 4.18. This behaviour leads us to believe that movements to and from Verona may not represent actual movements and that Verona should be interpreted carefully. However, this behaviour is unlikely to affect the spatial analysis of the next Chapter since the trips to and from stations outside of the Lombardia region (Verona Porta Nuova and Peschiera del Garda) will be excluded from our analysis because of the impossibility of defining their attracted municipalities

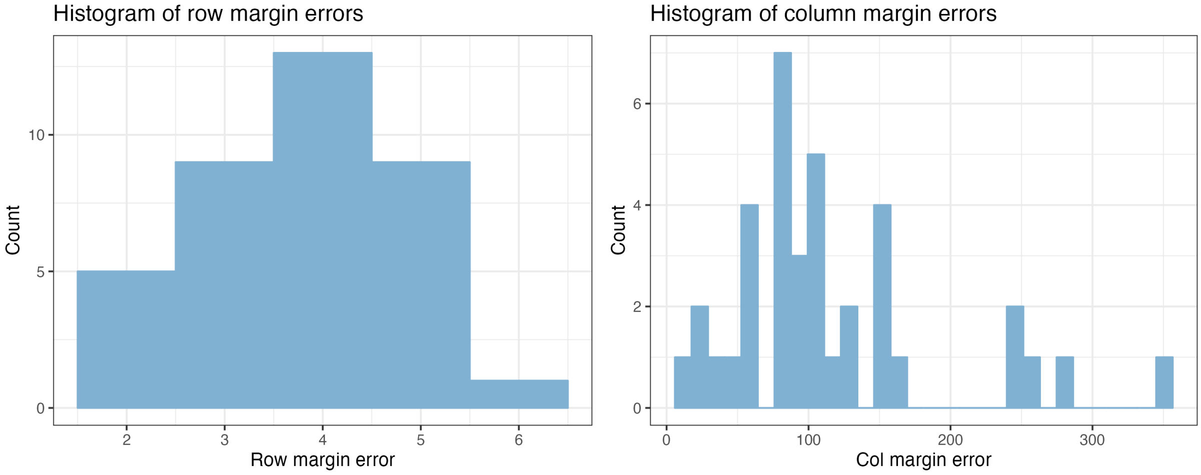


Figure 4.19: Histogram of row and column margin errors

(see subsection 4.4.1 for more details).

Another problem encountered during the application is the allocation of train journeys between two stations not connected by a direct line, which require a train change. For example, a journey from Monza to Seriate requires a train change, probably at Bergamo station. In this case, the ticket purchased for the journey shows Monza as the origin station and Seriate as the destination station. Still, the data from the passenger counters describe the journey as one passenger boarding at Monza, one passenger disembarking at Bergamo, one passenger boarding another train at Bergamo, and one passenger disembarking at Seriate, resulting in a total of two passengers boarding and two passengers disembarking. Due to this difference in the registration of the journey in the two datasets of tickets and passenger counts, there will be undefined behaviour in the registration of journeys with changes in the estimated OD matrices. Some of these journeys will be reported as direct journeys (Monza-Seriate in our example), derived from the correction of the ticket seed matrix. Others (we believe the majority) will be reported as separate journeys (Monza-Bergamo and Bergamo-Seriate), causing an underestimation of direct journeys with changes on the direct route. We have discussed possible solutions to this problem and concluded that reporting the direct journey in the OD matrices would be challenging due to the difficulty in identifying separate journeys in the counter data. It might be feasible to disaggregate the tickets into different paths, reporting each journey with a change as separate journeys in the matrices. However, one problem would be to identify the station where the change is made accurately: for some routes, it is easy (from Monza to Seriate, the change is most likely made at Bergamo), but for other routes, there are multiple possible changes (for example, from Rovato to Levate, the change could be made at Bergamo or Treviglio stations). The good news is that journeys with changes are likely a minority on networks such as ours, which cover long distances. Still, on short-distance networks such as the suburban network around Milan, the problem could be significant and require a deeper discussion.

4.4. Comparison between Regione Lombardia and Trenord OD matrices

After estimating the 37 Trenord OD matrices describing movements by train in a limited portion of the railway network, we want to investigate the relationship between the projected RL (described and explored in subsection 3.1.1) and the real Trenord data. To do so, we first need to restrict the RL data to the railway transportation mode. We have reason to believe that the majority of movements by train in the area happen through the Trenord network, as few other public railway services are available in the area (some exceptions are trips between major stations like Milan and Brescia, which could happen through some train lines owned by Trenitalia or Italo). After that, we have to define a common spatial granularity between the two datasets. Indeed, the Trenord OD matrices represent trips starting and ending from each station of the six train lines involved in the study. Thus, we have to define the area of influence of each station to assign the correct areal epidemic feature in the next Chapter's analysis. We aim to show differences and similarities in the two datasets, which we will refer to in the next Chapter when we will employ the two mobility datasets to analyse the relationship between mobility flows and the epidemic indicator.

4.4.1. Aggregation of municipalities into station's basins

To define the area of influence of each station, which we will refer to as "station's basin" going forward, we must determine the closest station for each municipality in Lombardia. To accomplish this, we assign each town to a single station. While we recognise that individuals residing in the same town may be closer to different stations depending on their precise location of residence, we are constrained by the granularity of the epidemic data, which is only available at the municipal level.

Thus, we need to perform an aggregation of stations belonging to the same municipalities. There are two cases of cities having multiple stations in their territory, Bergamo (with stations Bergamo and Bergamo Ospedale) and Treviglio (with stations Treviglio and Treviglio Ovest). Movements starting from and ending in these stations are aggregated into the two municipalities Bergamo and Treviglio, to obtain a correspondence with death counts data. Another case would be the city of Milan, but its stations have already been aggregated into the IS area in subsection 4.1.2.

To assign each municipality to the nearest station, we employ the ISTAT road distance dataset, described in subsection 3.1.3. This dataset provides information on the drive distance in meters and minutes between each pair of Italian towns (from one city hall to the other). Thus, we can only consider the distances between the city halls, which may significantly differ from the distance between each city hall and the station. We assign each municipality in Lombardia to the closest one having a Trenord station in its territory.

Furthermore, we encounter another limitation in this framework: while we have data on every station in Lombardia, we do not have data on all the stations in the Veneto region since the Trenord network only covers a limited number of stations in Veneto. Due to this, we decide not to define the station's basins for the two Veneto stations (Peschiera del Garda and Verona Porta Nuova). We thus restrict our analysis to the Lombardia region.

Despite these limitations, we determine the closest station for every municipality in Lombardia by identifying the station located in the town with the shortest road distance in meters to the target one. We have to exclude the town of Campione d'Italia from our analysis because the ISTAT road distance dataset contains no information about its distances to other cities. Additionally, the municipality of Monte Isola poses a challenge as it is situated on an island and lacks data on road distances. We were able to derive this missing information by calculating the distance by boat (5200 meters, 22 minutes [76]) to reach the closest municipality on land, Sulzano, and then compute all the other distances by adding the travel by boat to the road distances to and from Sulzano.

We select only municipalities whose closest station is one of the 42 stations in our study area after excluding the two Veneto stations and aggregating Bergamo and Treviglio stations. We further aggregate municipalities whose closer station belongs to one of the 15 stations of the IS area defined in subsection 4.1.2. Lastly, we exclude towns whose travel time to the nearest station is greater than 30 minutes since we evaluate that people travelling a considerable distance by car to get to the station are much more likely to use the car for the whole trip instead of taking the train.

After the whole procedure, our study area (which we abbreviate as BreBeMi area for brevity, keeping in mind that this area does not include the entire provinces of Milan, Brescia and Bergamo and includes some municipalities of the provinces of Lecco and Monza Brianza) consists of 261 towns referring to 28 station's basins. The area has a population of 3,294,217, a third of the Lombardia population. Figure 4.20 reports a map of the area. Each station basin is identified by the name of the municipality containing

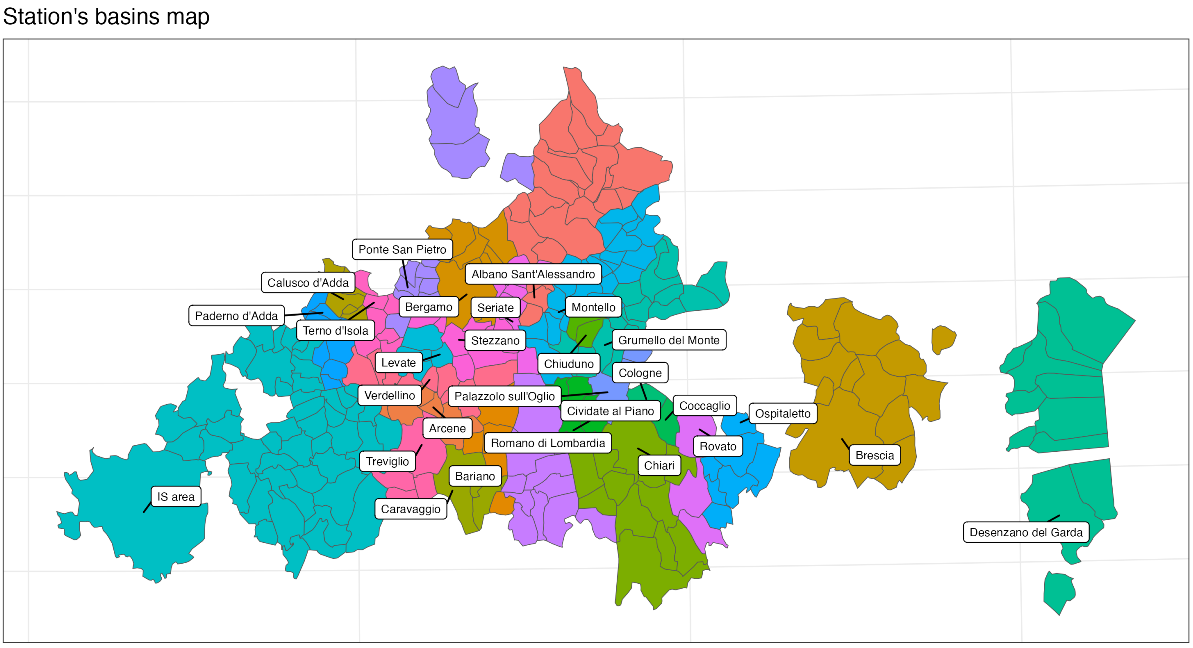


Figure 4.20: Station's basins map

the Trenord station (or stations) of reference, except for the aggregated IS area, identified by the term "IS area".

Finally, when aggregating the RL mobility data at the station's basins level, we deal with another problem: The RL dataset has four areas derived from an aggregation of municipalities where one of them has a station of the study area as station of reference and the others do not. Because of the limited number of railway movements starting and ending in these areas, we decide to refer the whole aggregate area to said station since it would not significantly affect our analysis. The problematic areas are *Casale Cremasco - Vidolasco - Castel Gabbiano*, with Castel Gabbiano having Morengo-Bariano as its station of reference, *Cortenuova - Parlasco - Taceno*, with Cortenuova having station Romano as its reference, *Parzanica - Tavernola Bergamasca*, with Tavernola Bergamasca having station Grumello, and *Val Brembilla - San Giovanni Bianco* with San Giovanni Bianco having station Ponte S. Pietro.

4.4.2. Comparison of mobility data through linear regression models

Our goal is now to compare the static RL OD matrix (for railway mobility only) with the corresponding dynamic OD matrix of Trenord, week by week. We exclude the four partial

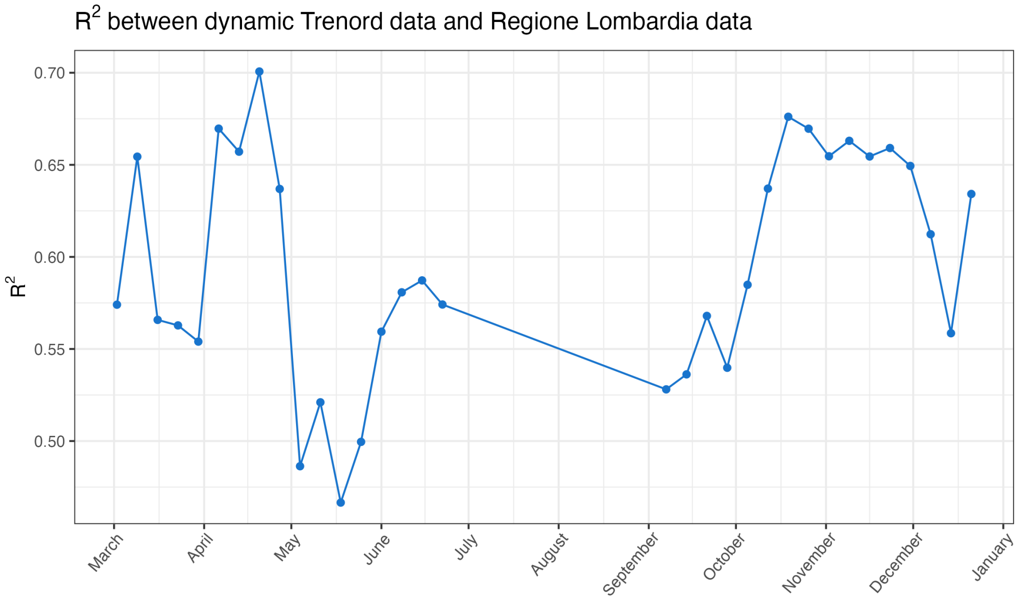


Figure 4.21: R^2 of the regression models built for comparing the mobility data

weeks of Trenord data (weeks 08, 26, 35 and 52) because of their incomplete representation of mobility trends.

In this framework, we build 33 regression models (one for every Trenord OD matrix) with the OD couples of the Trenord matrices as responses and the same OD couples of the RL dataset as unique feature without intercepts. We analyse the R^2 and the coefficients of the regression models.

When comparing the two kinds of mobility data, we must remember the different methodologies for estimating them. The Trenord OD matrices represent single trips by train in the area during the study period. We underline again the problem in allocating the trips requiring a change of train explained in subsection 4.3.3, which often causes single trips to be represented as more than one movement in the matrices. On the contrary, the RL OD matrix allocates a single trip for journeys having multiple means of transport, considering the start of the journey as the origin, the end as the destination, and the prevalent means of transportation. Thus, if a passenger takes the train but it is not the dominant means of transport for the trip, the movement will not appear in the railway RL OD matrix. Moreover, the railway transportation mode in the RL OD matrix reports data also about other transportation methods other than trains, like streetcars. Another criticality is that if a trip has origin (or destination) in the area of the study but destination (or origin) outside of it, it does not appear in the processed railway RL OD matrix even if one of the train lines involved in the study would be used for part of the trip. Finally, we must consider that there are weeks when some Trenord stations do not experience any movements (the inactive stations of Figure 4.8). In these cases, we remove the station's basins both for the Trenord and RL data because comparing them is misleading, as these are exceptional situations and do not reflect mobility changes.

R^2 coefficient Figure 4.21 represents the R^2 coefficients in the regression models. The average value is 0.59, which shows the power of the RL mobility data in predicting the Trenord movements. The coefficient varies in time during 2020. Particularly, R^2 is relatively low just after the beginning of the first lockdown in the middle of March and after the end of the lockdown in May, while it spikes in the period around Easter. During June, R^2 increases. Considering the second part of the year, R^2 increases again in September and October and stabilises during October and November, reaching values greater than 0.65. This hints at the fact that the projected RL OD matrix maintains a predictive power for real movements after the mobility disruptions happened in 2020. More effort should be put into investigating the mobility trends after the COVID-19 outbreak and establishing if the RL OD matrix still retains some validity in its mobility description.

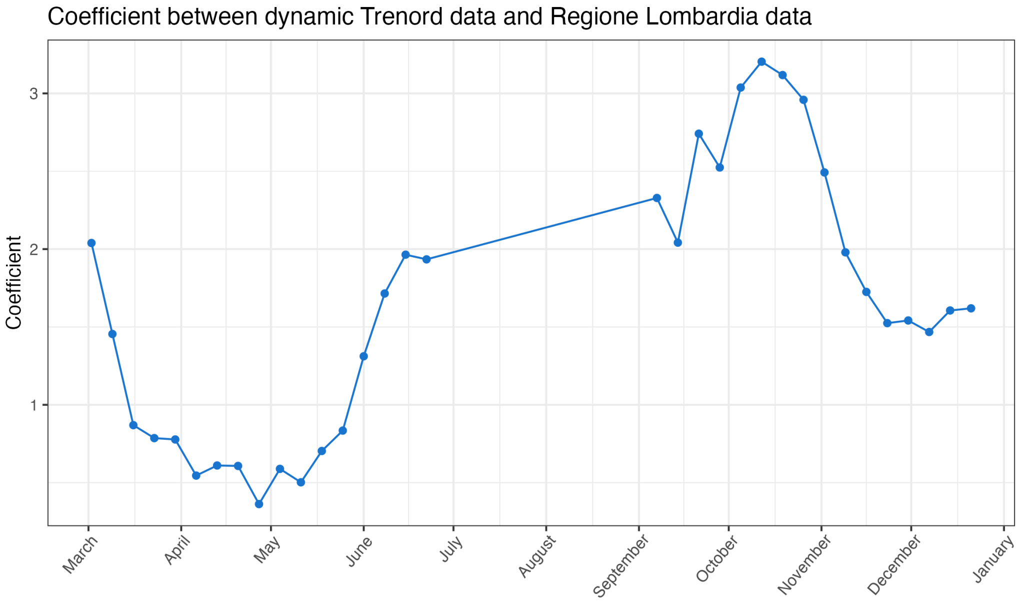


Figure 4.22: Coefficient of the RL mobility feature in regression models built for comparing the mobility data

Regression coefficients Figure 4.22 displays the estimated coefficients of the RL mobility feature in the regression models. Since the RL OD matrix represents railway movements of a single workday and the Trenord OD matrices refer to weeks, we would expect a coefficient greater than 5 between the RL data and the Trenord ones if mobility happened as expected. This never happens, probably because of the mobility changes due to the COVID-19 outbreak starting in March, after which railway mobility never resumed its full volume. However, we can observe a drop in the coefficient after the first week of March. The coefficient remains under the value of 1 for the first lockdown period, leading us to believe that movements by trains during this period were less than 20% of the expected movements. Mobility by train starts to increase after the beginning of June (sometime after the end of the lockdown), and the coefficient reaches a value close to 2 in the last week of June, which is still very far from the supposed pre-lockdown levels. In September and October, we see much higher values for the coefficient as mobility increases again after the Summer period, before the drop in values caused by the second wave and consequent second lockdown in Lombardia starting at the beginning of November. The mobility restrictions during the second lockdown were much less severe than those of the first one. Indeed, the coefficient during the second lockdown assumes values around 1.5, double those during the first lockdown. During December, when Lombardia was assigned to the orange and then yellow zone in the three-tier system, we see a slight increase in mobility following the end of the red-zone period.

Despite the significant differences in the two mobility descriptions, our analysis shows a strong relationship between them. The results indicate that the RL OD matrix, which represents railway movements of a single workday, has good predictive power for Trenord movements, especially in the last part of the year. However, due to the COVID-19 outbreak in early March, the regression coefficients between the RL data and the Trenord ones indicate reductions in mobility during the first and, to a lesser extent, the second lockdown. It could be fascinating to explore how the static projected RL OD matrix relates to the dynamic Trenord OD matrices describing mobility patterns in January and February before the pandemic outbreak. We have no data of this kind at the time, but if we could collect it, we could determine if there is a stronger correlation between the RL and Trenord data, indicated by higher values of R^2. We would also expect higher values of the regression coefficients in this period since the mobility volume was undoubtedly higher than in the rest of 2020.

To summarise, this analysis highlights the potential power of introducing the Trenord-derived OD data in the analysis of next Chapter since these OD matrices provide a detailed dynamic description of mobility through 2020. Indeed, they show the evolution of railway mobility flows closely related to the pandemic waves period and restrictions put in place because of the virus mentioned in Chapter 1.

4.5. Discussion

This Chapter developed a procedure to estimate dynamic OD matrices describing movements by train in a limited portion of the railway network.

Based on the Furness method, we built a general pipeline to reach the goal. The procedure could be enlarged in the future to include the whole Trenord network or applied to other kinds of transportation networks like metros, bus systems or other train services. The main data we need to estimate the OD matrices are OD data about survey-estimated trips in the network (in the form of tickets, turnstiles or tolls) and data about the number of trips starting and ending in each node of the network for each time frame.

We mentioned in Section 4.1 that our procedure presents some innovations compared to other works in trip distribution modelling. We now retrace the four steps of the pipeline and justify these innovations' contribution, supported by our results and by other analyses not shown in Section 4.3:

1. **Conversion of ticket data into seed OD matrices:** The ticket dataset reports data about different kinds of tickets and subscriptions, each with its own rules. We defined some assumptions to convert each ticket type into one or more ticket-estimated OD trips and derive ticket-estimated OD matrices, to be used as seeds in the Furness method. The 37 seed OD matrices obtained from this procedure show concordance with the margin vectors representing the number of boarded and dropped passengers at each station each week, according to the Wald test described in [66]. However, the estimation of seed OD matrices is not essential to apply the Furness method: [66] shows how the Furness algorithm could estimate OD matrices without the seed, using a seed matrix of cells $t^*_{ij} = 1 \ \forall(i, j)$. We tried this approach to assess the contribution of the ticket conversion step in the estimated OD matrices. We observed that the estimated OD matrices $T^{[w]}$ derived in the artificial seed case do not experience more movements than average between couples of stations near each other or connected by one or more direct train lines, opposite to what it is plausible to expect. The contribution of the tickets conversion step should be investigated more deeply, together with a sensitivity analysis to assess the robustness of the final OD matrices after Furness to the parametric tickets' assumptions. Still, we believe that this step strongly contributes to obtain OD matrices closely describing real railway mobility.

2. **Estimation of missing counter data:** We developed a model to correct the number of total boarded and dropped passengers for each station and each week of the

study, accounting for missing counter data. The model combines a linear regression with estimated trips from ticket OD matrices and the number of trains stopping in the stations as predictors and a rescaling procedure. It shows some desirable properties, never estimating negative values, limiting the number of underpredictions (i.e., predictions lower than the partial passengers' number from the aggregation of valid counter data) and estimating the counts for stations having no counter data for an extended period, showing similar trends to the stations having accurate counter data. Some issues in the predictions remain: five stations show underpredictions and nine stations present a peak in estimated passengers at the beginning of the first lockdown when the regression model plugs in, which we suspect does not describe the actual passengers' trend due to the number of estimated trips from tickets not happened because of the lockdown. However, the model represents the optimal outcome achievable given the available resources and constraints. Again, the Wald tests confirm the validity of our estimation process, highlighting the agreement between the ticket OD matrices and the margin vectors of boarded and dropped passengers for all weeks.

3. **Aggregation of the IS area:** Because we do not possess data about all the train lines moving in the area and because of the absence of a considerable number of tickets internal to the city of Milan, we aggregated the 15 stations of the IS area into a single zone. This choice is taken primarily because of the impossibility of defining accurate mobility flows for the area, which would influence the spatial analysis of the epidemic phenomenon presented in the next Chapter. If we had data about all the train lines moving in the area, we would have kept the stations separated and proceeded with the Furness method. We experimented with producing the estimated OD matrices while keeping the IS area disaggregated and came to sensible outcomes, showing similar properties to the OD matrices estimated aggregating the area. Still, we decided not to employ these matrices in the spatial analysis because of the sure underestimation of movements by train in the area.

4. **Application of Furness method:** Finally, we applied the Furness method to the 37 ticket-estimated seed OD matrices and margin vectors of boarded and dropped passengers. We used a procedure to correct the cells having no estimated trips from the ticket translation process. Indeed, because of the substantial underestimation in the ticket dataset, an OD couple might have a 0 in the ticket-estimated OD matrix, but trips have happened on the path. Since the Furness method cannot correct 0 values, we developed a procedure based on a posteriori binomial test to identify cells showing evidence of no trips happening between the couple while fixing with

positive values the other cells. Applying this procedure, we obtained matrices with lower margin errors than without the correction of zero values. Moreover, thanks to this procedure, we could estimate trip counts starting and ending in station Verona Porta Nuova since the corresponding entries in the seed matrices are all zero due to the missingness of ticket data.

Finally, we obtained 37 estimated OD matrices representing trips that happened in 37 weeks of 2020 between 32 sites of the Trenord network (31 stations and an aggregated area representing the 15 stations of the IS area). The final estimated OD matrices after Furness show coherency with some reality-induced principles, displaying many movements around major centres and on paths belonging to the same train line, revealing decreasing trips during the two lockdowns and increases in periods of lesser restrictions. The row margin errors, which measure the concordance between the estimated matrices and the margin vectors, were low (in the order of units) for all the matrices. Moreover, the iterative method always reached convergence. In the absence of objects describing actual mobility to compare with the derived OD matrices, the qualitative properties observed are the only possible pieces of evidence showing us that the final OD matrices could actually describe real railway mobility.

The main estimation issues concern station Verona Porta Nuova, caused by the absence of tickets to and from this station and the estimation of trips requiring a change of train discussed in subsection 4.3.3. While the first problem may be resolved by acquiring or estimating data about the missing tickets, the second one has no obvious solution and should be discussed more deeply in the case of future applications of the same pipeline.

After defining a common spatial granularity in the form of station's basins, we compared the Trenord dynamic OD matrices with the static RL matrix restricted to railway mobility through linear regression models. Results show a strong relationship between the two kinds of mobility data: the RL data have strong predictive power for the Trenord data, higher in periods of no mobility restrictions, and the regression coefficients through time show periods of lower mobility (during lockdowns) and periods of recovery, with railway mobility never quite reaching the expected level before the COVID-19 outbreak. We could interpret the similarity between the RL and Trenord-derived OD matrices as evidence of the pipeline's validity in estimating the Trenord matrices. Indeed, the two datasets are closely related and more so in periods of no mobility restrictions.

This analysis also suggests the potential retaining of some validity of the RL projected data after 2020 because of the predictive power shown for actual movements. We perform the comparison only for the limited portion of Lombardia, referring to the six train lines of the Trenord network and for railway transportation. However, we suggest further investigation into the relationship of the RL data with mobility after the pandemic.

5 | Spatial Analysis of Mobility and Epidemics

This Chapter develops a pipeline to analyse the relationship between an areal epidemic feature and a spatial description based on mobility flows, focusing on the role of railway mobility in 2020.

More in detail, Section 5.1 recalls some theoretical notions of spatial data analysis, defining spatial weights and the role of global and local Moran indexes in assessing spatial autocorrelation. Then, the Section presents the spatial analysis pipeline we developed to analyse two areas (the Lombardia and the BreBeMi areas) considering various mobility datasets. Section 5.2 applies the pipeline to the Lombardia area, considering the role of overall mobility (i.e., mobility considering all reasons and means of transportation) in the epidemic spread. Section 5.3 repeats the spatial analysis in the limited area and spatial granularity induced by Trenord data, considering the static projected RL OD matrix and the Trenord estimated OD matrices derived in Chapter 4 to adress the role of public railway mobility and to dynamically describe mobility trends after the COVID-19 outbreak. Finally, Section 5.4 discuess the results and answers the two research questions presented in the Introduction of this thesis.

5.1. Methodology

The analysis developed in this Chapter is based on the theoretical framework of spatial data analysis, a branch of statistics that deals with data analysis with a geographic or spatial component. This data type often has unique patterns and relationships that cannot be discovered through traditional statistical methods. A crucial aspect of spatial data analysis is the definition of spatial weights, which quantify the relationship between each data point and its neighbours. We can then assess spatial autocorrelation by computing Moran's index, both in its global and local form [77, 78]. The global Moran's index measures the overall spatial autocorrelation of a variable, meaning the degree to which the values of a variable at one location are related to the values of the same feature in nearby areas. On the other hand, local Moran's index measures the spatial autocorrelation of a variable at each location. For a deeper understanding of these concepts, see [79, Chapter 9] and [80].

In this Section, we define an areal epidemic indicator (subsection 5.1.1) and a spatial description based on mobility flows (subsection 5.1.2). Then we explain how to assess global and local spatial autocorrelation through global and local Moran indexes presented respectively in subsection 5.1.3 and subsection 5.1.4.

After presenting the necessary theoretical notions, subsection 5.1.5 explains the procedure adopted in analysing the two areas and the various mobility datasets employed in the analysis.

5.1.1. Definition of an epidemic indicator

To perform the spatial analysis, we have to define an areal dynamic epidemic indicator revealing the pandemic impact in each area. This epidemic indicator should not depend on the area's population and should adequately reflect fluctuations in mortality. For these reasons, and based on the ISTAT death data described in Section 3.2, we define the mortality index $m_i^{[w]}$ for each week $w \in [0, 52]$ of 2020 and each area of the spatial dataset considered $i \in I$ as:

$$m_i^{[w]} = \frac{\sum_{q=w-slide+1}^{q=w} deaths_i^{[q]}}{population_i} \tag{5.1}$$

The variable $m_i^{[w]}$ represents the sum of deaths of people aged over 70 years old in area i between the current week w and a specific number of weeks before that week, $w - slide + 1$. We always consider the area's population as of January 1, 2020, publicly available from ISTAT [59]. Moreover, we must determine an appropriate value for the parameter *slide* to ensure that $m_i^{[w]}$ does not oscillate too much while still reflecting weekly fluctuations in the mortality rate.

Handling the first and last weeks of the year with care is essential: week 0 includes the final partial week (two days) of 2019, while week 52 contains the first three days of 2021. Moreover, to calculate $m_0^{[w]}$, we must sum the death counts of weeks $[-1, 0]$, where week $w = -1$ is the second-to-last week of 2019 (since the final week is included in week 0).

We take the opportunity to define here a useful notation applied through the rest of the analysis: variables indicated with an apex in square brackets are variables calculated in the time instance indicated in the square brackets, while the pedex refers to the area in which the index is computed. The first example is $m_i^{[w]}$, the mortality index computed in area i at week w.

5.1.2. Spatial weights matrices

Spatial weights quantify the relationship between each data point in a spatial dataset and its neighbouring data points. To create the spatial weights associated with a dataset, we must first define which relationships between observations are given a non-zero weight and then assign weights to the identified neighbour links. The following steps of the spatial analysis will highly depend on the weights structure. The object obtained from the definition of spatial weights is a matrix δ, with each element of the matrix δ_{ij} indicating the strength of the relationship between two data points. Particularly, element δ_{ij} of the matrix is the weight of area i towards area j in the computation of spatial statistics. When $\delta_{ij} = 0$, area j is not a neighbour of area i. The relationship between data points is usually symmetrical ($\delta_{ij} = \delta_{ji}$), but nothing prevents us from defining asymmetrical spatial relationships. We define two descriptions of spatial relationships in our analysis, one based on contiguity and the other on a spatial description derived from mobility. We adopt the notation δ for the spatial weights matrix and δ_{ij} for single spatial weights, as opposed to the traditional weight notation W and w_{ij}. This choice has been taken because we use the letter w to indicate the time (in weeks) at which every value will be computed in the analysis and we wish to avoid notation confusion.

Contiguity-based spatial weights The first definition of spatial weights is based on a purely geographical description and corresponds to the most common one. Two areas are contiguous if they share a common border; in this case, we set $\delta_{ij} = 1$. Contiguity can be further distinguished between a rook and a queen criterion, in analogy to the moves allowed for the such-named pieces on a chess board. Thus, the rook criterion defines neighbours by having a shared edge between two spatial units, while the queen criterion defines neighbours as spatial units sharing a common boundary or vertex. Therefore, the neighbours' number for the queen criterion is always at least as large as in the rook criterion. We adopt the queen criterion in our analysis, thus defining the following

$$\delta_{ij} = \begin{cases} 1 & \text{if area } i \text{ and area } j \text{ share an edge or border} \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Mobility-based spatial weights The second description of spatial relationships between two data points considers the mobility flows between the areas. Suppose we have an OD matrix describing, for each ordered couple of areas (i, j), the number of movements

t_{ij} from i to j during a specific period. Then, we can define spatial weights in said period as

$$\delta_{ij} = \frac{t_{ij}}{\sum_{j \neq i} t_{ij}} \tag{5.3}$$

The spatial weights are row-standardised, dividing t_{ij} for the total number of outgoing trips from i, $\sum_{j \neq i} t_{ij}$. Notice that this definition is asymmetrical since usually $t_{ij} \neq t_{ji}$. This kind of spatial weights, as opposed to contiguity-based spatial weights, has rarely been applied in past studies. An example can be found in [50], where spatial weights are defined from mobility flows similarly and used to build a spatial econometric model for the diffusion of COVID-19 in China.

5.1.3. Global spatial autocorrelation

After defining a spatial weight structure as a matrix δ, we are ready to test for spatial autocorrelation. The term spatial autocorrelation refers to the presence of systematic spatial variation in a variable. A positive spatial autocorrelation indicates that adjacent observations (where "adjacent" refers to the notion of neighbouring described by the spatial weight structure) have similar values. In contrast, where adjacent observations tend to have contrasting values, we have negative spatial autocorrelation [81].

Moran's index [77] is one of the most commonly used indicators for spatial autocorrelation. It is a cross-product statistic between a variable m and its spatial lag δm, where the variable is expressed in deviations from the mean $m_i - \bar{m}$ for observation in location i. In our case, the univariate features will always be the epidemic indicator in the form of the mortality index defined in the previous subsection.

Given a spatial weights structure δ, Moran's I statistic can be computed as

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij}(m_i - \bar{m})(m_j - \bar{m})}{\sum_{k=1}^{n}(m_k - \bar{m})^2} \tag{5.4}$$

where $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij}$. Spatial weights are often row-standardised, as the mobility-based spatial weights presented in subsection 5.1.2, where $\sum_{j=1}^{n} \delta_{ij} = 1 \ \forall i$ such that $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} = \sum_{i=1}^{n} 1 = n$, where n is the number of observations in the spatial dataset. In this case, Moran's I simplifies to

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij}(m_i - \bar{m})(m_j - \bar{m})}{\sum_{k=1}^{n}(m_k - \bar{m})^2} \tag{5.5}$$

Moran's I ranges in the interval $[-1, 1]$. Values greater than 0 indicate evidence of positive spatial autocorrelation in the data, values around 0 point to spatial randomness and negative values are evidence of negative spatial autocorrelation.

Inference for Moran's I is based on the null hypothesis of spatial randomness. Tests rely on the assumption that the mean model removes systematic spatial patterning from data. Moreover, the spatial weights used to compute the statistic should adequately describe the process generating spatial autocorrelation. When testing, we always apply a permutational approach. The null hypothesis is expressed in the tests as

$$H_0 : I \leq 0 \qquad \text{vs} \qquad H_1 : I > 0$$

Our analysis sets the significance level at 0.05, adjusting the p-values for multiple testing when needed through the False Discovery Rate procedure [74].

5.1.4. Local spatial autocorrelation

The notion of global spatial autocorrelation presented in the subsection above describes the global relationship between the values observed at a spatial location and its neighbours, expressed by matrix δ. When the global tests reveal a significantly positive spatial autocorrelation, they indicate a clustered spatial pattern. However, this clustered spatial pattern does not point to the location of the clusters. To investigate this matter, the global relationship can be broken down into its components to build localised tests to detect *spatial clusters* - observations with very similar neighbours - and *spatial outliers* - observations with very different neighbours.

In this framework, Anselin [78] defines the concept of local indicator of spatial association (LISA). A LISA provides a statistic (together with its significance assessment) for each location and establishes a proportional relationship between the sum of the local statistics and the corresponding global statistics. Thus, a LISA can be derived from a global spatial autocorrelation statistic. A global spatial autocorrelation statistic consists of a combination of a measure of attribute similarity between a pair of observations $f(m_i, m_j)$ and an indicator of spatial similarity δ_{ij} in the form

$$\sum_i \sum_j \delta_{ij} f(m_i, m_j)$$

The corresponding LISA for location i has the form

$$\sum_j \delta_{ij} f(m_i, m_j)$$

As a result, the LISA counterpart of Moran's I at location i can be determined using the following equation

$$I_i = \frac{n}{S_0} \frac{(m_i - \bar{m}) \sum_{j=1}^{n} \delta_{ij}(m_j - \bar{m})}{\sum_{k=1}^{n}(m_k - \bar{m})^2}$$

If row-standardised weights ($S_0 = n$) are used, the formula simplifies as in the case of global Moran's statistics

$$I_i = \frac{(m_i - \bar{m}) \sum_{j=1}^{n} \delta_{ij}(m_j - \bar{m})}{\sum_{i=1}^{n}(m_i - \bar{m})^2}$$

However, in our application, we use the modified version of these formulas presented in [67], which considers that when we calculate I_i for area i, we use data from the other $n - 1$ observations. This modified formula is as follows

$$I_i = \frac{n-1}{S_0} \frac{(m_i - \bar{m}) \sum_{j=1}^{n} \delta_{ij}(m_j - \bar{m})}{\sum_{k=1}^{n}(m_k - \bar{m})^2} \tag{5.6}$$

The simplified formula for row-standardised weights is given by

$$I_i = \frac{n-1}{n} \frac{\sum_{j=1}^{n} \delta_{ij}(m_i - \bar{m})(m_j - \bar{m})}{\sum_{i=1}^{n}(m_i - \bar{m})^2}$$

Once again, we have to assume that the global mean $\bar{m}$ adequately represents the variable of interest m.

The local Moran's I_i's value at location i identifies i as part of a *spatial cluster* if I_i assumes positive values. On the contrary, location i is a *spatial outlier* if I_i is negative. Significance can be tested again using methods based on permutations. In this case, it is critical to overcome the normality assumption when computing significance because the number of neighbours for each observation can be minimal, which may make adopting the normality assumption problematic. The null hypothesis of the local tests is two-sided since we are interested in detecting both positive and negative values:

$$H_0 : I_i = 0 \qquad \text{vs} \qquad H_1 : I_i \neq 0$$

Assessing significance in and of itself is not that useful for the local Moran. The local indication of significance is usually combined with the location of each observation in the Moran scatterplot [82]. This plot represents the original variable with the mean removed $(m_i - \bar{m})$ on the x-axis and the spatially lagged variable $\sum_{j=i}^{n} \delta_{ij}(m_j - \bar{m})$ on the y-axis. The slope of the linear fit to the scatterplot equals global Moran's I. Indeed, the simplified formula in Equation (5.5) turns out to be the slope of a regression of $\sum_{j=1}^{n} \delta_{ij}(m_j - \bar{m})$ on $(m_i - \bar{m})$. Figure 5.1 shows an example of a Moran scatterplot. This plot allows us to classify the nature of spatial autocorrelation into four categories. Since the graph is centred on the mean, all points with values smaller than the mean $\bar{m}$ have negative values on the x-axis and positive values if they are greater than $\bar{m}$. We refer to these values respectively as *high* and *low*, in the sense of higher and lower than average. Similarly, we can classify the values for the spatial lag above and below the mean as *high* and *low*. The scatterplot is then decomposed into four quadrants. The upper-right and lower-left quadrants correspond with positive spatial autocorrelation (similar values at neighbouring locations). We refer to them as respectively *high-high* and *low-low* spatial autocorrelation. In contrast, the lower-right and upper-left quadrants correspond to negative spatial autocorrelation (dissimilar values at neighbouring locations). We refer to them as respectively *high-low* and *low-high* spatial autocorrelation.

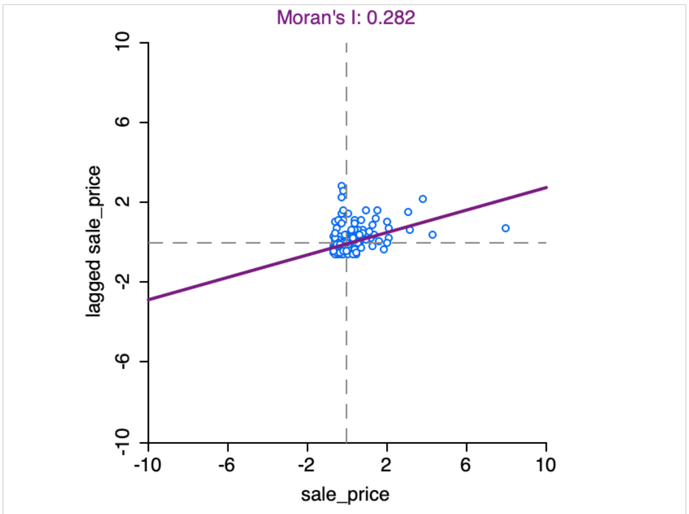


Figure 5.1: Example of a Moran scatterplot for variable `sale_price`, taken from [80]

The classification of the spatial autocorrelation into four types can then be interpreted together with the significance of each location's local Moran's I_i. Significant locations are *high-high* and *low-low* spatial clusters and *high-low* and *low-high* spatial outliers. In a spatial cluster, the feature m_i has a value which deviates (positively for *high-high* or negatively for *low-low*) from the overall mean in the same direction as the average behaviour of the neighbours of i. On the contrary, if location i is a spatial outlier, m_i deviates from the overall mean in the opposite direction as the average neighbours. Again, our analysis considers a significance level of 0.05, adjusting for multiple testing when needed.

5.1.5. Spatial analysis pipeline

In the next Sections of this Chapter, we aim to analyse global and local spatial autocorrelation in a spatial univariate epidemic feature through spatial descriptions inferred from mobility flows.

To reach our goal, we select a region divided into areas $i \in I$ and choose a time period divided into weeks $w \in W$. We then define the following steps to assess spatial autocorrelation in an epidemic feature through mobility-based spatial descriptions:

1. We define an appropriate areal dynamic epidemic indicator $m_i^{[w]}$ based on mortality data.

2. We infer a spatial description based on mobility flows in the form of spatial weights matrices δ_{ij}. This spatial description can be static or dynamic in time. Its definition is based on mobility data in the form of OD matrices. We also show a classical spatial description in the form of contiguity spatial weights, which we will compare to the mobility-based one in our analysis. When considering dynamic mobility-based spatial weights, we have to explore possible lag values between the epidemic indicator and the corresponding dynamic spatial weights.

3. We compute weekly global Moran indexes $I^{[w]}$, $w \in W$, and test for positive global spatial autocorrelation in the epidemic feature. Positive values of the indexes reveal a clustered tendency in the epidemic indicator, with nearby areas experiencing similar values. The underlying hypothesis is that no positive spatial autocorrelation would be revealed in the absence of an epidemic phenomenon.

4. Lastly, we compute weekly local Moran indexes $I_i^{[w]}$, $w \in W$, to detect spatial clusters and outliers, i.e., observations with nearby areas having very similar or dissimilar values of the epidemic indicator. In particular, spatial clusters with higher-than-average values of the epidemic indicator reveal epidemic hotspots (areas of elevated disease burden). The evolution in time of these areas could be studied through the dynamic index.

This procedure is graphically summarised in Figure 5.2.

We apply the procedure to two regions:

1. **Lombardia area:** First, we conduct the spatial analysis on the entire Lombardia region. We define mobility-based spatial weights using the RL OD matrix described in subsection 3.1.1. This dataset covers a large region in fine spatial granularity. For this reason, it is the first mobility dataset considered in our analysis. How-

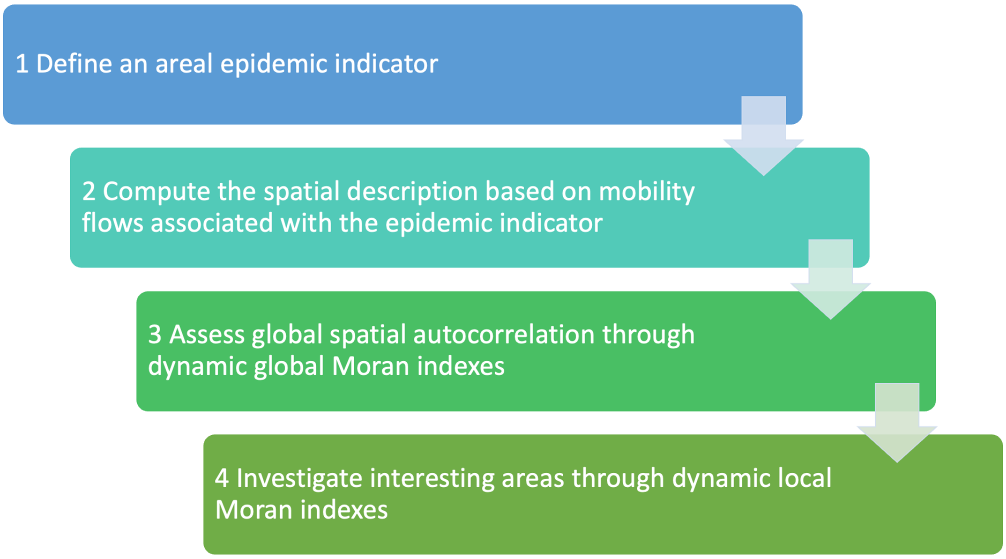

1 Define an areal epidemic indicator
2 Compute the spatial description based on mobility flows associated with the epidemic indicator
3 Assess global spatial autocorrelation through dynamic global Moran indexes
4 Investigate interesting areas through dynamic local Moran indexes

Figure 5.2: Procedure to assess the spatial relationship between mobility flows and an epidemic feature

ever, the dataset is static as it is derived from projections of data collected prior to 2020. Thus, it does not represent actual mobility in 2020 and we can hypothesise that its description loses explanatory power after the mobility disruption following the COVID-19 outbreak. We will apply the spatial analysis pipeline to the mobility flows in the dataset considering overall mobility. Then, we compare the results obtained with the same analysis considering spatial weights derived from a geographical description based on contiguity.

2. **BreBeMi area:** After the first level of analysis, we want to deepen our understanding of the relationship between mobility and epidemic spread through dynamic mobility representations. In this framework, we employ the weekly Trenord OD matrices derived in Chapter 4. These matrices cover the area that we will abbreviate with the name BreBeMi from now on for brevity since most of its territory belongs to the provinces of Brescia, Bergamo and Milano, even if some municipalities in the area find themselves in the Lecco and Monza Brianza provinces. The spatial granularity is the station's basins defined in subsection 4.4.1. In this case, we compare overall mobility with railway mobility to address the role of a particular kind of public transport believed to be a primary carrier of epidemic diffusion and strongly impacted by restrictions. Notably, we consider two mobility datasets:

    (a) *RL static OD matrix:* First, we retrace the analysis of the RL dataset in the new area and spatial granularity to compare the results with the ones obtained in the Lombardia area. We also compare mobility flows derived from overall and railway mobility to introduce the role of railway mobility in the analysis and later compare it with the Trenord data

(b) *Trenord dynamic OD matrices:* We add dynamic mobility-based spatial weights describing railway mobility in the analysis. The description of actual railway mobility overcomes the limitations of the projected RL mobility data, allowing us to have a reliable description of the mobility changes in 2020 after the COVID-19 outbreak. We can also assess in detail the role of public railway transport. In this case, we have to explore lag values (in weeks) between the mobility-based spatial weights and the epidemic indicator.

Figure 5.3 summarises the analysis of this Chapter. The light-blue boxes show the considered regions (Lombardia and BreBeMi), the green ones specify the mobility datasets chosen for each area (together with the contiguity description adopted in the Lombardia area), and the blue ones display the type of mobility considered (overall or railway mobility) at each level.

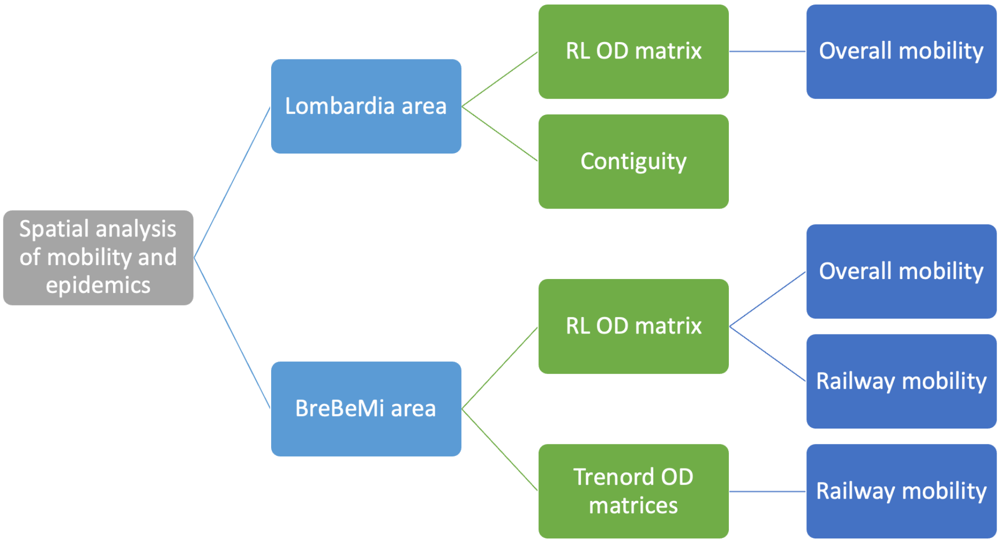


Figure 5.3: Representation of the two areas considered in the spatial analysis, together with the spatial mobility description adopted.

We can now proceed to apply the spatial analysis pipeline to all the areas and mobility datasets explained in this subsection.

5.2. Spatial analysis of the Lombardia area

This Section introduces the first application of the spatial analysis pipeline between epidemic and mobility data. We first focus on the Lombardia area to exploit the mobility description of the RL dataset.

More in detail, subsection 5.2.1 explores the weekly mortality rates in the area through 2020 and subsection 5.2.2 investigates spatial association by comparing a spatial description based on the contiguity of investigated zones with the one derived from the static projected RL mobility data. We first assess the presence of global spatial autocorrelation and then detect interesting locations from an epidemic point of view.

The main goal of this Section is to establish whether there exists a link between the evolution of the pandemic in 2020 and mobility.

5.2.1. Exploratory analysis of ISTAT death data

Based on the ISTAT death data described in Section 3.2, Figure 5.4 illustrates the weekly death counts of people aged over 70 years old in Lombardia's provinces. The curves' evolution is closely related to the pandemic's spread. All provinces show a peak in death counts during the disease's initial months, around May and April, with Brescia and Bergamo experiencing the sharpest rises. We can observe a second, less pronounced peak during the second wave in November. This peak is more flattened for those provinces strongly hit during the first period of the epidemic.

The population size of different areas heavily influences death counts. To account for this impact, we defined the mortality rate expressed in Equation (5.1), selecting $slide = 2$ to aggregate death counts in time. We justify the choice of this parameter in Appendix A for reasons inferred from the spatial analysis results. Furthermore, to align the death counts dataset with the OD matrix that represents projected movements for 2020 in Lombardia (described in subsection 3.1.1), we have to aggregate some municipalities to obtain matching spatial granularity. Specifically, we aggregate the 1504 municipalities in Lombardia to recover the 1360 zones of the RL OD matrix.

We compute the mortality index $m_i^{[w]}$ for every week of 2020 ($w \in [0, 52]$) and every area identified by the OD matrix spatial granularity ($i \in [1, 1360]$). Figure 5.5 displays the mortality rate $m_i^{[w]}$ computed in six weeks after the spatial aggregation:

- As expected, we can observe the impact of COVID-19 during the first wave in weeks 10 and 12 in March and April. During this period, we can notice a clear spatial

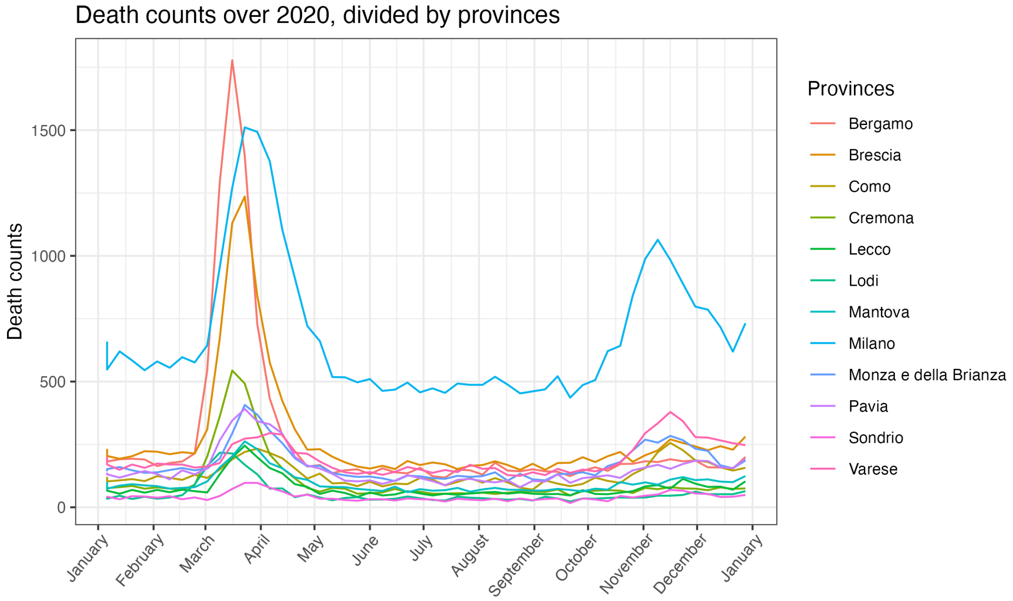


Figure 5.4: Death counts of people aged 70+ years during 2020, divided by provinces

trend in the mortality rate, with high values of $m_i^{[w]}$ in the areas surrounding Val Seriana and Codogno in week 10 and an increase in $m_i^{[w]}$ in the Brescia province starting from week 12.

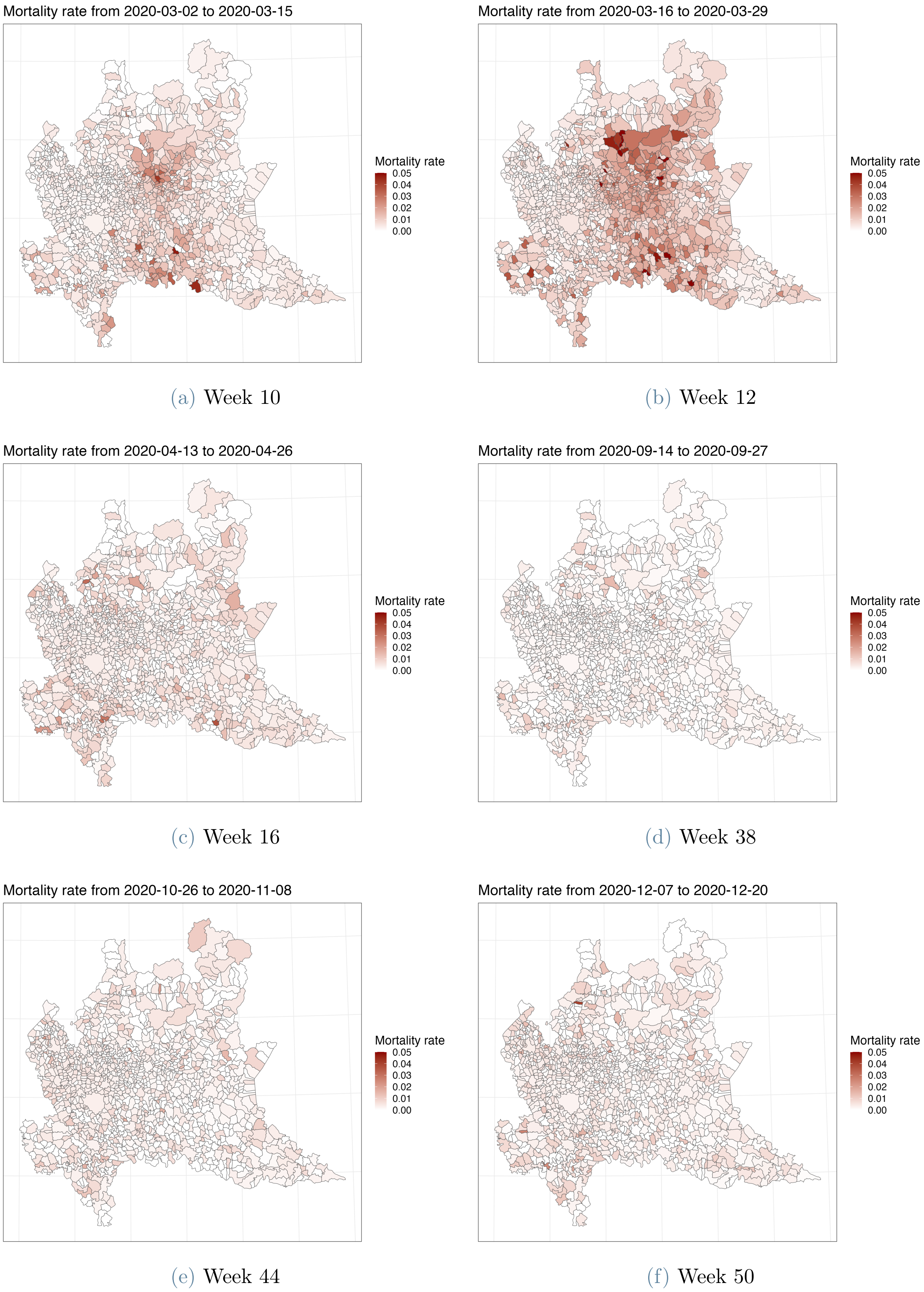
- Following the lockdown, which halted non-essential movements in Lombardia from March 8 until May 18, the mortality index generally decreases throughout Lombardia, as seen in week 16.

- During the second half of the year, week 38, which falls in September before the beginning of the second wave, displays a generally low mortality rate throughout Lombardia, with no evident hotspots.

- Week 44, during the second wave, shows a slight general increase in the mortality rate but no areas with values much higher than the average, indicating homogeneity in the disease spread.

- Finally, week 50 shows the reduction in mortality following the introduction of the three-tier system, which was implemented to restrict mobility during the second wave. Lombardia was designated a red zone, with maximum restrictions, from November 6 until November 29. This is the second lockdown period in the region.

This exploratory analysis allows us to confirm that there are two periods (that we can identify with the first and second waves of the epidemic) where the mortality index $m_i^{[w]}$ generally increased and then decreased sometime later, following the adoption of restrictions and lockdowns. The first wave displays a diffusion of the epidemic starting from some specific hotspots. In contrast, the epidemic follows a more spatially-homogeneous diffusion during the second wave.


Mortality rate from 2020-03-02 to 2020-03-15
Mortality rate from 2020-03-16 to 2020-03-29
Mortality rate from 2020-04-13 to 2020-04-26
Mortality rate from 2020-09-14 to 2020-09-27
Mortality rate from 2020-10-26 to 2020-11-08
Mortality rate from 2020-12-07 to 2020-12-20
Mortality rate
0.05
0.04
0.03
0.02
0.01
0.00

(a) Week 10

(b) Week 12

(c) Week 16

(d) Week 38

(e) Week 44

(f) Week 50

Figure 5.5: Mortality index $m_i^{[w]}$ in Lombardia during some weeks of 2020

5.2.2. Spatial analysis of Regione Lombardia mobility data

We will now investigate the relationship between the epidemic indicator $m_i^{[w]}$ defined in subsection 5.1.1 and the spatial description derived from the RL mobility dataset in the Lombardia area, applying the spatial analysis pipeline shown in Figure 5.2.

We define the two spatial weights detailed in subsection 5.1.2 using Equation (5.2) and Equation (5.3), denoted as δ_C and δ_{OD}, respectively. The former is based on contiguity, while the latter is based on mobility, considering the overall movements in Lombardia while disregarding internal area movements. There is a relationship between the two definitions, as a significant proportion of movements starting from a zone tends to be directed towards contiguous areas. However, mobility-based weights expand the spatial description beyond contiguity and allow us to define a novel notion of nearness. According to this new notion, two municipalities can be tightly related despite the distance between them (think, for example, about the Milan metropolitan area, where many people travel a considerable distance from and to Milan for work or study reasons every day).

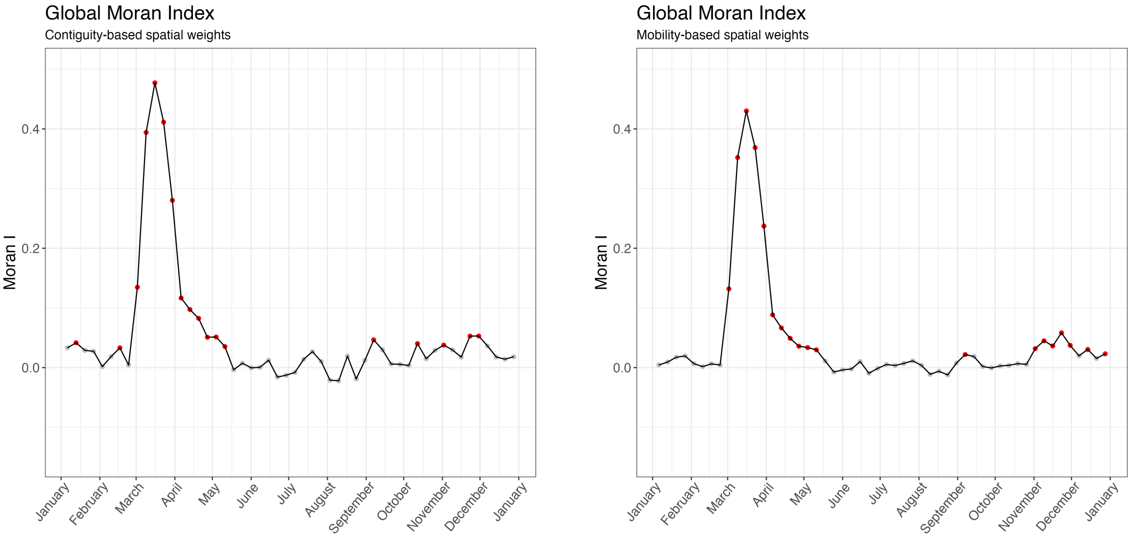
We compute global and local Moran indexes of the mortality index $m_i^{[w]}$ for each week of 2020 ($w \in [0, 52]$) and each of the 1360 areas under investigation ($i \in [1, 1360]$). Moran indexes allow us to explore the relationship between mobility and the epidemic at a granular level and shed light on potential spatial patterns and possible clusters. In this case, the dynamicity of the analysis derives from the mortality index $m_i^{[w]}$, which evolves during the year, while the spatial weights remain fixed. In Section 5.3, we will see another case of the same analysis where the spatial weights vary in time.

Global Moran indexes

We compute global Moran indexes $I^{[w]}$ for each week of 2020, $w \in [0, 52]$, using Equation (5.4) and the methodology described in subsection 5.1.3. We test for positive spatial autocorrelation of the epidemic indicator $m_i^{[w]}$ for each week $w \in [0, 52]$ using both contiguity-based and mobility-based spatial weights. The municipality Campione d'Italia is excluded from the contiguity-based analysis due to a lack of shared borders or vertices. Still, we can define its neighbours using mobility-based weights. Moreover, we adopt the same aggregation of municipalities into 1360 zones induced by the RL OD matrix when computing the contiguity-based spatial weights to compare maps having the same spatial granularity using the two spatial descriptions.

We compute the curves that display the global Moran indexes' evolution for different values of the *slide* variable used to calculate the mortality indexes $m_i^{[w]}$ in Equation (5.1).

We select *slide* = 2 for reasons justified in Appendix A that we suggest reading after the end of this Section. From now on, we always consider *slide* = 2 in the computation of $m_i^{[w]}$.



(a) Contiguity-based spatial weights

(b) Mobility-based spatial weights

Figure 5.6: Global Moran indexes at Lombardia level based on RL mobility data

Figure 5.6 compares the global Moran indexes computed for each week and spatial weight description. The red points on the graph indicate a significantly positive spatial autocorrelation for the epidemic phenomenon described by $m_i^{[w]}$ in a given week (i.e., adjusted p-values smaller than 0.05). The plot reveals several significant findings:

- During the epidemic outbreak period (March) and immediately after, the epidemic phenomenon is strongly dependent on both spatial descriptions, with positive spatial autocorrelation detected from week 9 (beginning on March 2) until week 19 (ending on May 17).
- After May, both spatial descriptions do not reveal significantly positive spatial autocorrelation in the mortality indexes. Notably, this coincides with the end of the national lockdown on May 18, suggesting the effect of the substantial restrictions during the lockdown period in removing the spatial pattern of mortality.
- Comparing the two spatial weights δ_C and δ_{OD}, we observe that the corresponding Moran indexes are significantly greater than 0 in the same weeks during the first wave. Both weights are almost equivalent in describing the epidemic phenomenon, but the Moran indexes induced by contiguity weights are slightly greater than those generated by mobility weights.

- We do not detect spatial autocorrelation during the summer months by any investigated spatial description. In general, we found some weeks showing positive spatial autocorrelation outside the two wave periods but never extended in time. This confirms our initial hypothesis that no spatial autocorrelation is detected in the absence of an epidemic phenomenon.

- Starting from week 44 (the first week of November, beginning on November 2), we observe another period of positive spatial autocorrelation detected using the mobility weights δ_{OD} that continues until the end of the year, during the second wave period. We note a clear increase in the curve describing the evolution of the global Moran index, albeit more oscillating and less evident than the one observed during the first wave. This trend occurs despite the projected RL OD matrix no longer describing mobility trends after March. However, the spatial weights defined from the RL OD matrix significantly describe the epidemic phenomenon, indicating that the projected movements still provide valuable insights into the actual mobility patterns. This behaviour is consistent with the comparison between the static RL OD matrix and the dynamic Trenord OD matrices estimated in Chapter 4, where Section 4.4 highlighted the strong relationship that occurs between the projected static RL data and the dynamic Trenord ones estimated from data collected in 2020, suggesting that the RL data retains some predictive power for real mobility and more so in periods of no restrictions to limit the epidemic phenomenon. We suggest further investigation into the relationship between the RL OD matrix and actual movements in 2020.

- Considering contiguity weights δ_C, we also notice some weeks showing positive spatial autocorrelation towards the second wave period. However, the trend here is more oscillating and does not show a clear peak compared to the mobility weights δ_{OD}. Thus, mobility weights reveal more about the epidemiological phenomenon during the second wave than contiguity-based ones.

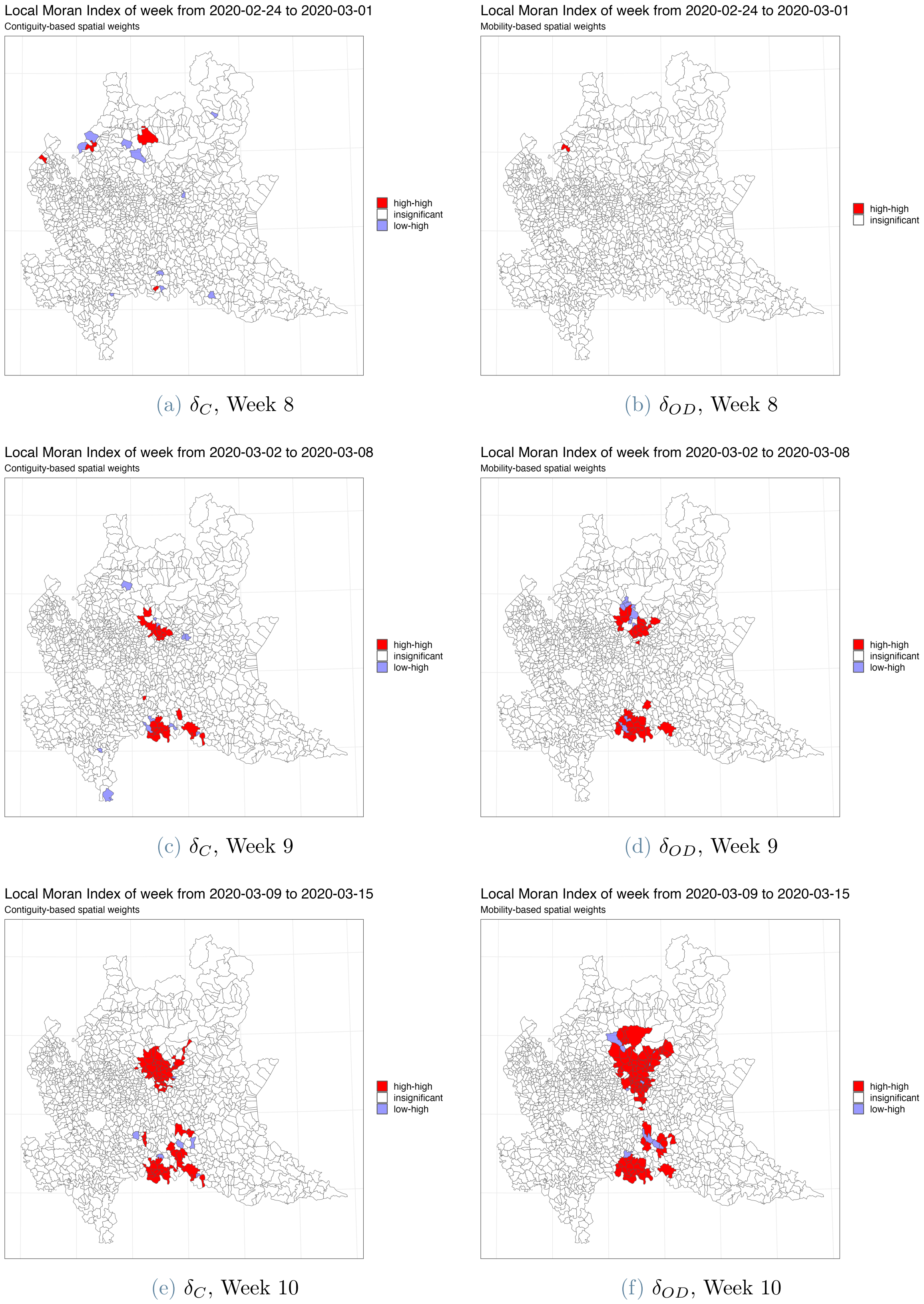
Local Moran indexes

After establishing the global trends of spatial autocorrelation in the epidemic phenomenon during 2020, we aim to identify interesting areas for each week. Starting from the theoretical framework described in subsection 5.1.4, we compute the local Moran indexes $I_i^{[w]}$ using Equation (5.6) for each week $w \in [0, 52]$ and each area $i \in [1, 1360]$ of the RL OD matrix. We assess the significance of each index using permutational tests and proceed to identify interesting areas.

We focus on identifying *high-high* spatial clusters because of their potential impact on the spread of the pandemic. The epidemic heavily affected these areas, which show mortality indexes higher than the average, with nearby areas also showing higher-than-average mortality indexes. We investigate the characteristics of these areas and their association with epidemic hotspots, as determined by two spatial descriptions: contiguity-based δ_C and mobility-based δ_{OD}. We highlight that the difference in the two spatial descriptions stands in the definition of neighbouring areas and the weighting of such areas in the computation of the spatially lagged values $\delta(m - \bar{m})$. Indeed, fixing an area i, weights δ_C identify as its neighbours the zones sharing a border or vertex with i and weights each zone equally. On the contrary, weights δ_{OD} do not limit the number of neighbours of i (defined with all the areas with at least one person moving there from i) but weights the importance of each neighbour according to the proportional mobility flow outgoing from i.

We underline that we could not conclude that one or the other spatial descriptions could identify epidemic sources in the sense of places where the epidemic originated its spread. We merely developed a tool to flag areas showing clustered values of the epidemic feature according to different notions of nearness. However, more analyses should be conducted to assess the role of the flagged areas in epidemic diffusion. We could elaborate on this matter by analysing the time series of the epidemic feature for nearby areas, where the term "nearby" has to be read together with the spatial description adopted. In particular, we would choose *high-high* areas (identified in the local spatial analysis) and nearby zones (looking for high values in the spatial weights) to identify anticipators in the first time series for the second one. This procedure could compare notions of nearness (mobility or contiguity-based, or even other notions not explored in this work) and assess if one of them is firmly related to the evolution of the epidemic indicator. However, previous studies discussed in Chapter 2 support a strong relationship between mobility and epidemic spread, which implies that the *high-high* areas flagged according to the mobility-based spatial description may have played a crucial role in the epidemic's propagation.

Coming to our results, both the spatial weights definitions δ_C and δ_{OD} reveal extensive *high-high* spatial clusters during the first wave. The spatial clusters are persistent in time and found in some areas identified by other works, such as [10], to have been severely hit by the epidemic. Figure 5.7 and Figure 5.8 show the evolution of the LISA maps from week 8 to week 13 (from February 24 to April 5, the period of the strongest epidemic spread). Each row of the figures depicts the same week $w \in [8, 13]$. The left column represents the indexes computed using contiguity-based spatial weights δ_C, while the right column represents the same indexes calculated with mobility-based spatial weights δ_{OD}.†


Local Moran Index of week from 2020-02-24 to 2020-03-01
Contiguity-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-02-24 to 2020-03-01
Mobility-based spatial weights
🟥 high-high
□ insignificant
Local Moran Index of week from 2020-03-02 to 2020-03-08
Contiguity-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-02 to 2020-03-08
Mobility-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Contiguity-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Mobility-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high

(a) δ_C, Week 8

(b) δ_{OD}, Week 8

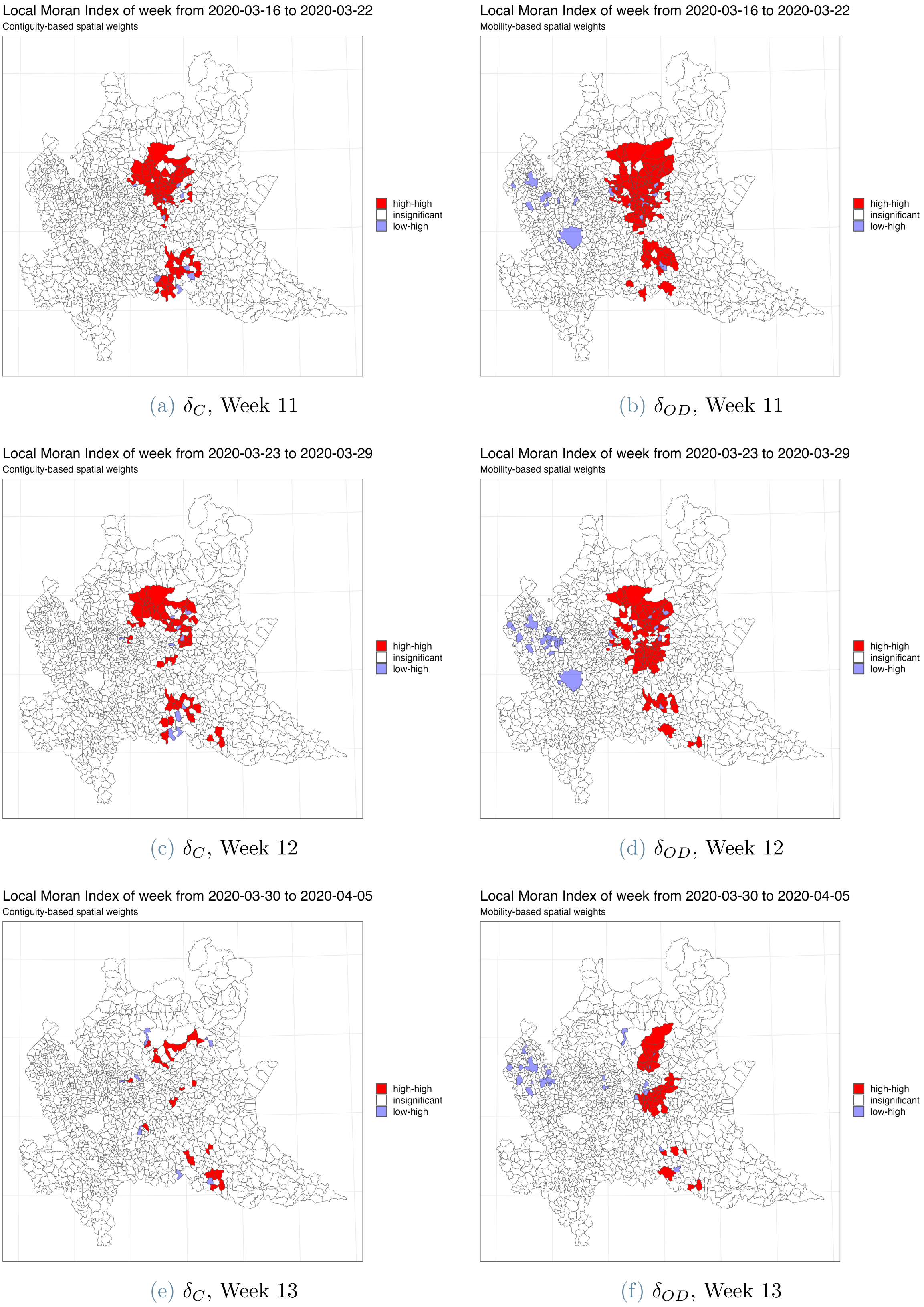
(c) δ_C, Week 9

(d) δ_{OD}, Week 9

(e) δ_C, Week 10

(f) δ_{OD}, Week 10

Figure 5.7: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia



Local Moran Index of week from 2020-03-16 to 2020-03-22
Contiguity-based spatial weights

(a) δ_C, Week 11

Local Moran Index of week from 2020-03-16 to 2020-03-22
Mobility-based spatial weights

(b) δ_{OD}, Week 11

Local Moran Index of week from 2020-03-23 to 2020-03-29
Contiguity-based spatial weights

(c) δ_C, Week 12

Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights

(d) δ_{OD}, Week 12

Local Moran Index of week from 2020-03-30 to 2020-04-05
Contiguity-based spatial weights

(e) δ_C, Week 13

Local Moran Index of week from 2020-03-30 to 2020-04-05
Mobility-based spatial weights

(f) δ_{OD}, Week 13

Figure 5.8: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia

The first thing we can notice is the evolution in time of the spatial clusters, looking at both columns. We do not see any effect of the epidemic in week 8, but starting from week 9, we notice two red areas appearing on the maps. We can recognise the northern area as the Val Seriana, containing Nembro and Alzano, and the southern area as Codogno and its surroundings. The latter was the first area where the first case of COVID-19 was found in Italy and the first area to be placed in the red zone on February 21 [2]. Thus, it makes sense that it would be the first area where the pandemic started to take its toll on mortality. It is more interesting to notice (even if already established by past studies) that the pandemic effect hit the mortality of the Val Seriana area at the same time as the Codogno area. A third red area arises in the Brescia province starting from week 10. This area expands in the following weeks, while we see the Codogno area gradually disappearing, showing the effect of the early restrictions applied there. From week 13 onwards, all the red areas shrink until their complete disappearance in the following weeks.

The analysis's second and most significant finding is that mobility-based spatial weights identify larger spatial clusters than contiguity-based weights through local Moran indexes. Indeed, throughout the studied period, except for week 8, the δ_{OD} approach consistently reveals larger red zones than the δ_C approach.

To further investigate the *high-high* areas detected by δ_{OD} but not δ_C, we examine the difference in the local Moran indexes $I_i^{[w]}$ values for areas with higher-than-average mortality rates. For example, the city of Bergamo belongs to a *high-high* spatial cluster for weeks 9 to 13 according to δ_{OD} but not according to δ_C. Thus, if we only consider the 14 towns sharing a border or vertex as Bergamo's neighbours, we do not reveal Bergamo as part of a spatial cluster. However, if we expand the number of Bergamo neighbours and weight each neighbour's contribution according to mobility flows, the area is part of a spatial cluster for an extended period.

As stated before, the consequence of the epidemic in Bergamo (and other areas similarly flagged as *high-high* through δ_{OD} but not according to δ_C) in neighbouring zones should be investigated in detail. If the hypothesis that the disease spread in nearby areas through mobility flows is confirmed, identifying high-high areas according to mobility-based spatial weights could provide policymakers with a system to identify sources of the epidemic phenomenon.

Considering the second wave period, we do not detect any area extended in space or time experiencing consistently higher-than-average mortality index. However, we have to remember the limited power of the RL OD matrix in describing mobility after the first months of 2020. Still, the global spatial analysis shows that there seem to be some insights in the projected data about an association between mobility and epidemic during the second wave. The fact that the contiguity-based spatial weights still fail to reveal *high-high* areas suggests that the epidemic phenomenon could be homogeneous during the second wave and no more characterised by hotspots. Coupling this finding with the global spatial association, we acknowledge that similar mortality index values still tend to be near each other, without particular locations showing a clustered spatial pattern.

To summarise the key findings of this Section, the analysis conducted in the Lombardia area demonstrates the potential of mobility data in understanding spatial patterns of mortality rates during the COVID-19 pandemic. The analysis found that mobility data can be just as effective as contiguity-based spatial weights in identifying global spatial autocorrelation during the first wave of the pandemic. In addition, mobility-based spatial weights detected an epidemic phenomenon during the second wave, showing more weeks with significantly positive spatial autocorrelation than contiguity-based spatial weights. The study also revealed that when examining local spatial patterns, mobility data consistently identified larger spatial clusters than contiguity-based methods. The findings suggest that further investigation into mobility data is needed, considering dynamic mobility data describing how mobility evolved throughout 2020 and its influence on the epidemic phenomenon. This need motivated the development of the procedure to derive dynamic OD matrices presented in Chapter 4.

5.3. Spatial analysis of the BreBeMi area

This Section reproduces the spatial analysis between epidemic and mobility data of Section 5.2, introducing the real mobility data derived in Chapter 4 to overcome the limitations of the projected Regione Lombardia data through dynamic OD matrices representing actual weekly railway mobility in 8 months of 2020. We analyse the BreBeMi region, induced by the available Trenord data, considering the spatial granularity defined by the station's basins derived in subsection 4.4.1.

More in detail, subsection 5.3.1 repeats the exploratory analysis of the mortality rates in the newly defined area and spatial granularity, pointing out the evolution of COVID-19 in the area during the pandemic's first and second wave periods. After this, subsection 5.3.2 reproduces the spatial analysis of the mortality rates using mobility-induced spatial weights from the RL data in the BreBeMi area, comparing overall mobility with railway mobility. This subsection is needed to connect Section 5.2 with the analysis of the Trenord OD matrices and introduce the study of railway mobility. Finally, subsection 5.3.3 plugs the real Trenord mobility data into the spatial analysis pipeline to confirm the previous findings through actual data and investigate the second wave period in detail, for which we previously had no reliable mobility description.

This Section elaborates on the results obtained in Section 5.2, deepening our understanding of the link between the evolution of the pandemic during 2020 and mobility and developing the analysis to address the role of public railway transport in the pandemic diffusion, compared to overall mobility.

5.3.1. Exploratory analysis of ISTAT death data at station's basins granularity

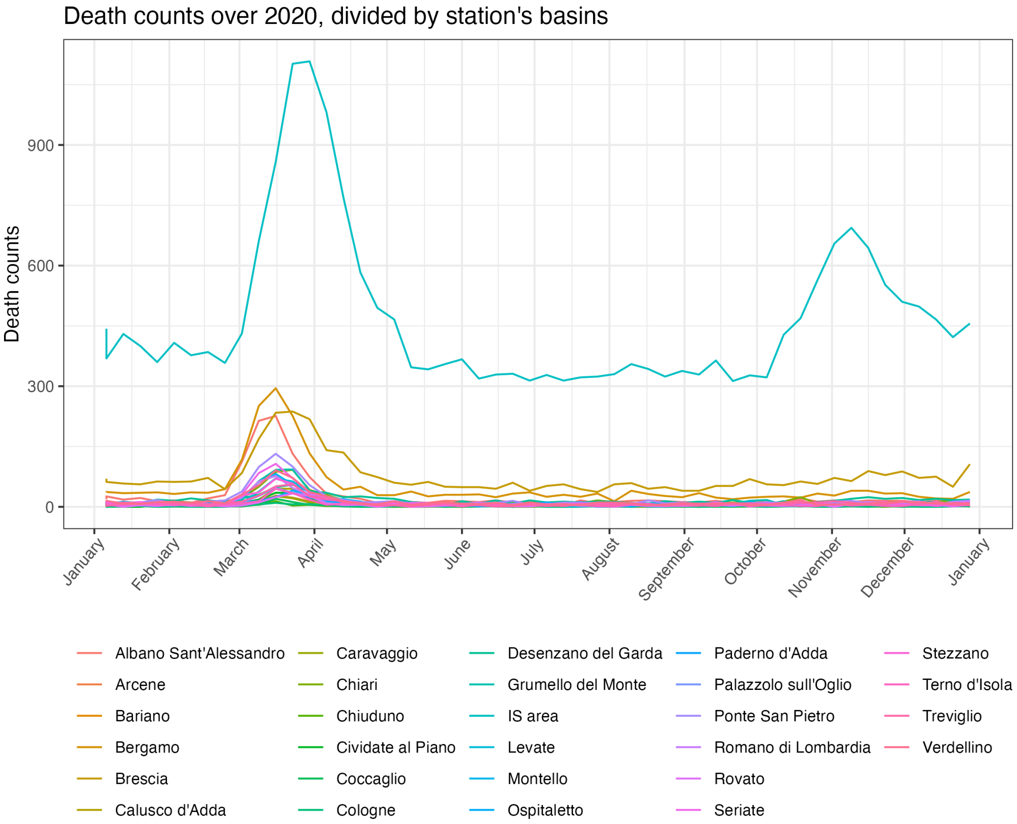
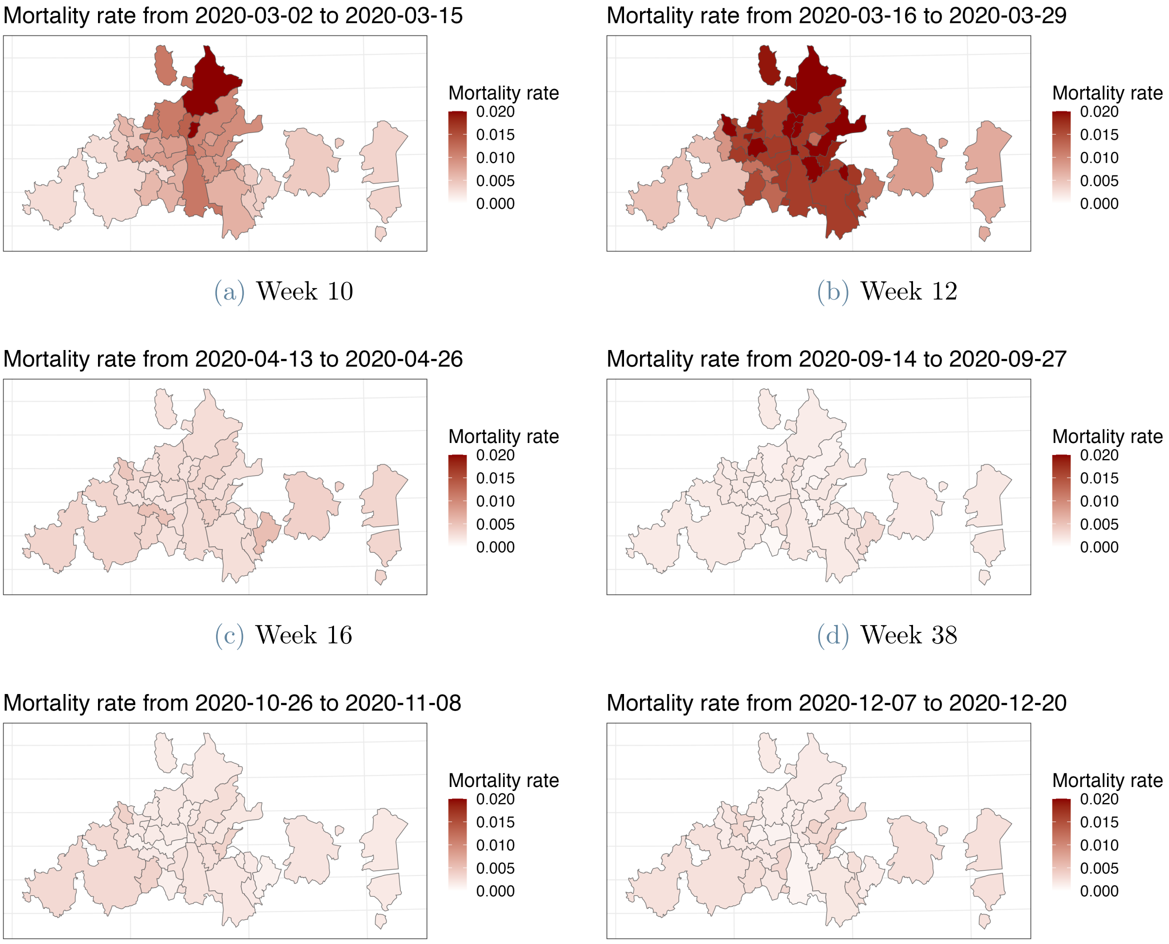


Figure 5.9: Death counts of people aged 70+ years during 2020, divided by station's basins

Figure 5.9 shows the weekly death counts of people aged over 70 years old in the BreBeMi area at the station's basins level. We can notice a peak in all the basins during the first wave period, with peaks happening at different times. Indeed, the mortality rise in the IS area happens some weeks later than the one in Bergamo. Basins Bergamo, Albano S. Alessandro, Brescia and IS area experience the most significant increases in mortality since they are the most populated basins. During the second wave, the figure allows us to notice only the peak of the IS area.


Mortality rate from 2020-03-02 to 2020-03-15
Mortality rate from 2020-03-16 to 2020-03-29
Mortality rate from 2020-04-13 to 2020-04-26
Mortality rate from 2020-09-14 to 2020-09-27
Mortality rate from 2020-10-26 to 2020-11-08
Mortality rate from 2020-12-07 to 2020-12-20
Mortality rate
0.020
0.015
0.010
0.005
0.000

(a) Week 10

(b) Week 12

(c) Week 16

(d) Week 38

(e) Week 44

(f) Week 50

Figure 5.10: Mortality index $m_i^{[w]}$ in the BreBeMi area at station's basins level during some weeks of 2020

We compute the value of the mortality index $m_i^{[w]}$ defined in Equation (5.1) for each of the 28 station's basins and each week w of 2020, $w \in [0, 52]$. We select again $slide = 2$ for a temporal aggregation of the death counts, as was done in Section 5.2. We consider the whole 52 weeks of year 2020 in the analysis of the RL data, while we select only 33 for the Trenord analysis, excluding the partial weeks in Trenord data shown in Figure 4.2.

Figure 5.10 displays the mortality rate $m_i^{[w]}$ computed for six different weeks at the station's basins level:

- As expected, we can observe the impact of COVID-19 during the first wave in weeks 10 and 12 in March and April. In particular, week 10 shows high values of $m_i^{[w]}$ for the basin of Albano S. Alessandro, which contains some municipalities of the Val Seriana area, a zone known to be strongly hit by the disease.

- After some weeks, we can notice COVID-19 spread in week 12 in the whole Bergamo province and also in some territories in the Brescia province, while the IS area containing the Milan province, Brescia basin containing municipalities around the city of Brescia and the Desenzano del Garda basin do not experience an equal mortality rise.
- Following the first lockdown, the mortality index generally decreases in the area, as seen in week 16.
- During the year's second half, week 38 displays a generally low mortality rate throughout the area, with no evident hotspots.
- Week 44, during the second wave, shows a slight general increase in the mortality rate. The increase is slightly more evident in the IS area around Milan.
- Finally, week 50 shows a small reduction in mortality following the second lockdown period, after the introduction of the three-tier system.

This exploratory analysis, similarly to the one in subsection 5.2.1, shows two periods (that we can identify with the first and second waves of the epidemic) where the mortality index $m_i^{[w]}$ generally increased and then decreased sometime later following the adoption of restrictions. The first wave displays a stronger epidemic diffusion in some specific locations, like Albano S. Alessandro basin.

In contrast, the epidemic follows a more spatially-homogeneous diffusion during the second wave. Compared with the same exploratory analysis performed in the Lombardia area, the second wave of the epidemic did not significantly affect the mortality rates in the BreBeMi area, except for the territories around Milan. [10] also notices this behaviour and hypothesises that the reason for the low mortality rates during the second wave compared to other zones in Lombardia is that the pandemic stroke the Bergamo and Brescia areas extremely strongly during the first wave. Thus, the second one was almost imperceptible, presumably because the population at high risk had already significantly shrunk, and survivors had developed natural immunity due to the large circulation of the virus during the first wave.

5.3.2. Spatial analysis of Regione Lombardia mobility data

We can now investigate the spatial relationship between the mortality index $m_i^{[w]}$ and the different mobility descriptions provided by the RL and Trenord data.

We first analyse the RL mobility data to reproduce the spatial analysis of Section 5.2 in a different area and granularity, corresponding to that induced by real Trenord mobility data. We then compare this analysis (based on projected data about overall and railway mobility) with the real railway mobility analysis of subsection 5.3.3.

We do not consider contiguity-based spatial weights in this analysis (as we did in Section 5.2) since the station's basins' granularity does not allow us to define such weights. Indeed, most areas do not have a complete representation of all their neighbours, and some areas (Brescia and Desenzano del Garda) have no neighbours at all. Instead, we compare two types of mobility through mobility-based spatial weights: overall mobility (i.e., mobility derived by the sum of single movements by all reasons and transportations) and railway mobility (i.e., mobility derived by the sum of single movements by all reasons, considering only railway transportation). We refer to the two spatial weights as δ_T for overall mobility and δ_R for railway mobility. We aim to compare the two mobility descriptions and assess their relationship differences with the epidemical phenomenon.

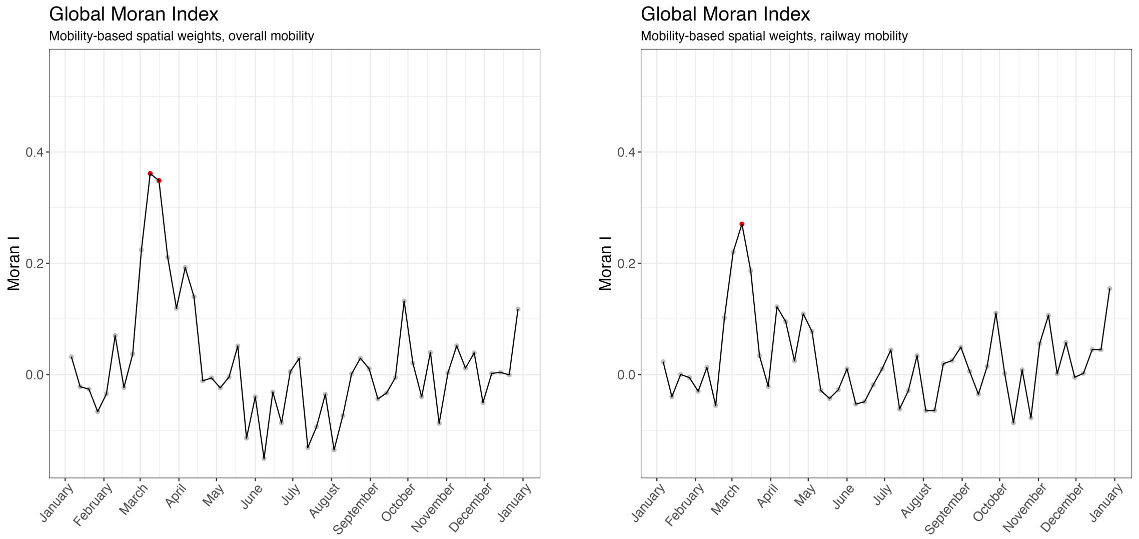
We compute global and local Moran indexes, following the pipeline described in subsection 5.1.5, considering the mortality index $m_i^{[w]}$ for each of the 28 areas and the 52 weeks of 2020. The two spatial weights, δ_T and δ_R, are fixed in time, while the index $m_i^{[w]}$ varies weekly. We can compare the results with those obtained in the Lombardia area at municipalities' granularity in subsection 5.2.2, in which case we considered contiguity-based spatial weights δ_C and overall mobility-based spatial weights δ_{OD}, equivalent now to δ_T.

Global Moran indexes

We compute global Moran indexes $I^{[w]}$ for each week $w \in [0, 52]$ of 2020, applying both the spatial weights based on overall mobility δ_T and those generated by railway mobility δ_R. We use the methodology described in subsection 5.1.3, testing for positive spatial autocorrelation in each index generated by the two mobility descriptions.

Figure 5.11 displays the global Moran indexes computed using the spatial weights δ_T and δ_R. Red points indicate a significantly positive spatial autocorrelation for the epidemic phenomenon described by $m_i^{[w]}$ in the given week (i.e., adjusted p-values smaller than 0.05). The graph allows us to comment on some points:

- First, we compare the overall mobility weights δ_T with the same weights δ_{OD} at the



(a) Overall mobility-based spatial weights δ_T

(b) Railway mobility-based spatial weights δ_R

Figure 5.11: Global Moran indexes at BreBeMi level based on RL mobility data

Lombardia level in Figure 5.6. We can notice the much more oscillating trend and the lower values the indexes assume at the BreBeMi level. There are many fewer weeks where we can identify positive spatial autocorrelation in the epidemic phenomenon. Regarding this, the limited area considered in this analysis (the BreBeMi area, in contrast with the whole Lombardia area) and the wider spatial granularity (station's basins, as opposed to municipalities) undoubtedly play a role. Because of the limited area, observing differences in the mortality indexes compared to the average values is much more challenging. Moreover, the small number of areas complicates assessing significance in the permutational tests.

- Nevertheless, it is still possible to sense a positive spatial autocorrelation during the first wave period. We find evidence of that in both the spatial descriptions δ_T and δ_{OD}. Thus, the epidemic phenomenon during the first wave is related to the mobility description of the RL dataset, which represents mobility in an average workday during the first months of 2020. We came to the same conclusion in the analysis conducted at the Lombardia level.

- Again, as in the Lombardia level analysis, the positive spatial autocorrelation disappears from data during the first lockdown period. We have evidence of the lockdown's effectiveness in breaking the connection between the epidemic spread and mobility. In the BreBeMi level case, the hypothesis of spatial randomness can not be refused starting from the last week of March (for weights δ_T). Positive spatial autocorrelation disappears much earlier than what we observe at the Lombardia level, but still, we have to underline the role of the limited area and wider spatial granularity in the analysis.

- In Section 5.2, we observed a period of positive spatial autocorrelation during the second wave at the Lombardia level, even though the projected mobility derived from the RL data is no more reliable after March 2020. In contrast, the second wave period shows no positive spatial autocorrelation at the BreBeMi level. The reason we do not observe the same behaviour here could be found in the exploratory mortality analysis of subsection 5.3.1, where we noticed the little impact that the second wave had on mortality in the territories of Bergamo and Brescia and the homogeneity in the mortality indexes during this period.
- Comparing overall mobility described by δ_T with railway mobility as δ_R, the two graphs show remarkably similar trends for the global Moran indexes $I^{[w]}$. The tests reveal two weeks of positive spatial autocorrelation for weights δ_T (weeks 10 and 11, in March) and only one week for δ_R (week 10). There are no weeks when δ_R weights show significantly positive spatial autocorrelation and δ_T weights do not, and this is the first hint that railway mobility probably did not play a prevalent role in the spread of the epidemic compared to overall mobility.

Local Moran indexes

While global Moran indexes allow us to identify periods of global trends of positive spatial autocorrelation, we rely on local Moran indexes to investigate interesting areas showing values of the mortality index $m_i^{[w]}$ similar or dissimilar to nearby locations. The term "nearby" refers to the notion of nearness derived from the spatial weights δ_T and δ_R. We compute the local Moran indexes $I_i^{[w]}$ for every week $w \in [0, 52]$ of 2020 and area identified by the station's basins $i \in [1, 28]$. The theoretical framework can be found in subsection 5.1.4. We assess the significance of each index using permutational tests.

We never reveal any significant area in the weeks of 2020 for each spatial description adopted. This is the opposite of our findings in the spatial analysis at the Lombardia level, which detected an extensive high-high spatial cluster in the Val Seriana area during the first wave of 2020. We can also fathom the same high mortality region in the exploratory mortality analysis in subsection 5.3.1. We can only guess two possible reasons for this behaviour:

1. The local spatial pattern of the epidemic spread may have no link with the two mobility descriptions. However, it is unlikely that the two phenomena are uncorrelated, as the analysis at the Lombardia level shows a strong relationship between

them, individuating extensive epidemic hotspots in the Val Seriana area.

2. The absence of significant areas in the local spatial analysis may be imputed to the extension and subdivision of the area considered. The variations of the mortality indexes compared to the average weekly values are minimal because of the limited extension of the area considered, where the indexes tend to rise and decrease together, and the small number of data areas (28 for each week). Moreover, even when we observe values of $m_i^{[w]}$ higher than average, assessing the significance of the local Moran index with respect to the other 27 areas through permutational tests is challenging. This is likely the reason for the difficulties in detecting spatial clusters and outliers at this level, which could only be overcome by expanding the analysis area. The spatial granularity can not be reduced because we need to assign every municipality to the station of reference to match the spatial analysis considering Trenord mobility data, which can not be reproduced otherwise.

5.3.3. Spatial analysis of Trenord mobility data

Finally, we reproduce the spatial analysis of subsection 5.3.2 (same area and spatial granularity), plugging in the real railway movements estimated in Chapter 4. The description of actual mobility overcomes the limitations of the projected RL mobility data, allowing us to have a reliable description of the mobility changes in 2020 after the COVID-19 outbreak. We can also assess in detail the role of public railway transport.

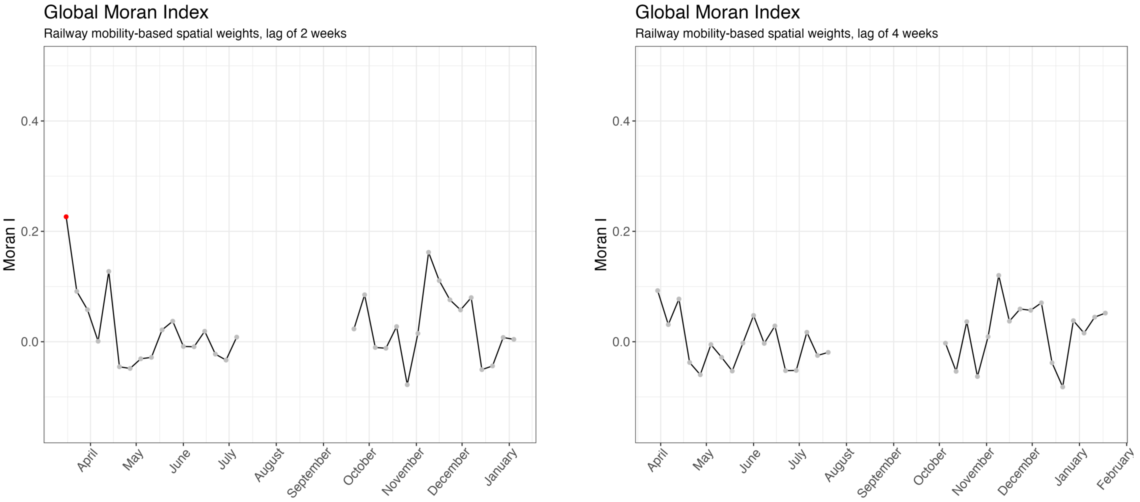
In Chapter 4, we estimated 37 weekly OD matrices representing train journeys in the study area during 8 months of 2020 (from March to June and September to December). We remove 4 of these weeks since they do not contain 7 days of data and use the remaining 33 OD matrices to build railway mobility-based spatial weights $\delta_R^{[w]}$ as described in subsection 5.1.2, applying again Equation (5.3). Unlike the other mobility-based spatial analysis of subsections 5.2.2 and 5.3.2, the mobility-based spatial weights $\delta_R^{[w]}$, $w \in W \backslash \{08, 26, 35, 52\}$ are dynamic and change weekly. Thus, we have two dynamic levels in this final spatial analysis: the dynamicity of the mortality index $m_i^{[w]}$ and the one of the spatial weights $\delta_R^{[w]}$.

We hypothesise a lag between the mobility phenomenon and the consequence on mortality rates, as was found in other works such as [11, 12]. We thus compute the global and local Moran indexes $I^{[w]}$ and $I_i^{[w]}$ selecting the mortality rates $m_i^{[w]}$ of the same week w and considering spatial weights $\delta_R^{[w-q]}$, where q is the lag in weeks between the mobility and epidemic effects. We vary $q \in [0, 10]$. Values of $[w - q]$ for $\delta_R^{[w-q]}$ can fall only in $W \backslash \{08, 26, 35, 52\}$, the set of weeks for which we have an estimated OD matrix from the

Trenord network (see Figure 4.2 for a visual representation of W, excluding weeks 08, 26, 35 and 52). We define mortality rates $m_i^{[w]}$ for 2021 when needed, encoded by $w > 52$. The inactive stations shown in Figure 4.8 must be ignored in the spatial analysis because of the impossibility of defining their mobility-based spatial weights.

Global Moran indexes

We compute global Moran indexes $I^{[w]}$ for each of the weeks $[w-q] \in W \backslash \{08, 26, 35, 52\}$ and each possible lag $q \in [0, 10]$. We find a single case in which we have evidence of positive spatial autocorrelation, corresponding to week of mortality index $w = 11$ and lag $q = 2$, associated with mobility weights $\delta_R^{[9]}$. Fixing the same mobility weights $\delta_R^{[9]}$ and varying the values of q to study the effect on other mortality rates, no influence of the same mobility weights on other weeks' mortality rates is found through global Moran indexes.



(a) $q = 2$

(b) $q = 4$

Figure 5.12: Global Moran indexes at BreBeMi level based on Trenord mobility data, comparing two values for the lag q to define railway mobility-based spatial weights $\delta_R^{[w-q]}$

Figure 5.12 shows the evolution of the global Moran indexes choosing two different values for the lag, $q = 2$ or $q = 4$. First, we want to compare the real railway mobility inducing spatial weights $\delta_R^{[w-q]}$ with the estimated railway mobility of the RL dataset inducing weights δ_R, whose global Moran indexes are shown in Figure 5.11. The Trenord-induced weights show more oscillations and lower values in the Moran indexes than the RL-induced ones.

Considering the first wave period, we have to underline that the first available Trenord estimated OD matrix refers to the first week of March, the only one found significant in the analysis (spatial weights $\delta_R^{[9]}$). This matrix describes the week from March 2 to March 8, when the first Italian cases of COVID-19 had already been announced, and fear of the virus was influencing people's movements. We thus have reason to believe that the influence of railway mobility on mortality rates during the first wave period should be investigated considering February mobility, before the announcement of the COVID-19 outbreak and the following mobility restrictions. Adding real mobility data for this period is undoubtedly an important future development of this work. Still, our study retains validity since we have no reason to believe that mobility in the first months of 2020 significantly differs from the projected RL data. In our analysis, we rely on the RL mobility data to describe railway mobility in February, and the Trenord data provides actual descriptions after that.

Putting together the two kinds of mobility data, we can conclude that there is a spatial association between railway mobility and the epidemic phenomenon during the first wave period. We find evidence of that in subsection 5.3.2, where we observe the influence of RL projected railway mobility on the mortality rates during the second week of March. Moreover, we also infer an effect of real railway mobility in the first week of March on the mortality rates of the third week in this subsection. However, when considering the role of railway mobility compared to overall mobility in the epidemic spread during the first wave, as highlighted in subsection 5.3.2, we do not observe any evidence in favour of a leading role in the diffusion of railway mobility compared to overall mobility. Indeed, we found no weeks (here and in subsection 5.3.2) where railway mobility correlates with the epidemic phenomenon and overall mobility contradicts this behaviour.

Coming to the analysis of the second wave period, we have no problem with missing mobility data here since the outbreak of the second wave can be pinned on the September and October period, and we have Trenord data from September to December. We must underline that these data are the only reliable ones to describe actual mobility during the second wave period in our study. We came to some conclusions using the projected RL data in subsections 5.2.2 and 5.3.2, but we know we can not trust the RL mobility description after March. The Trenord data finally allow us an accurate mobility description for this period.

No global Moran indexes show positive spatial autocorrelation in the second wave period: plugging in real mobility data into the pipeline did not change the findings of the RL data analysis in subsection 5.3.2. We remind again that the BreBeMi area, as shown in the exploratory analysis of subsection 5.3.1, was essentially untouched in mortality during the second wave period because of the strong impact experienced there during the first outbreak of the epidemic. This fact could explain the absence of positive spatial autocorrelation in the second wave period, while we observed a global effect in the Lombardia area considering the projected RL mobility data in subsection 5.2.2.

Local Moran indexes

We compute local Moran indexes $I_i^{[w]}$ for each lag $q \in [0, 10]$, each w *s.t.* $[w - q] \in W \backslash \{08, 26, 35, 52\}$ and each of the 28 station's basins. We aim to detect interesting locations showing values of the mortality index $m_i^{[w]}$ similar or dissimilar to nearby areas in the sense of railway mobility-based weights $\delta_R^{[w-q]}$.

Local Moran indexes $I_i^{[w]}$ never reveal any significant areas for any value of $q \in [0, 10]$ and w *s.t.* $[w - q] \in W \backslash \{08, 26, 35, 52\}$. We found the same behaviour in the analysis of the RL projected data at the BreBeMi level of subsection 5.3.2. We can thus confirm and expand the comments made for the RL mobility data case.

Considering the first wave period, we have to underline again that if a link between railway mobility and the disease spread existed, it would need to be thoroughly researched in the period where mobility happened without influence from events associated with the epidemic spread. However, we can rely on the RL data to describe that period (February) to complement the analysis. Despite the evident spatial pattern highlighted in subsection 5.3.1, where we can see high values of the mortality indexes in some basins, particularly in those belonging to the Bergamo province, the local spatial analysis reveals no significant areas. Again, the same two hypotheses described before could explain this behaviour. First, the absence of spatial outliers and clusters could be explained by the lack of connection between railway mobility-based spatial weights and local spatial patterns in the epidemic feature. Secondly, the limited extension and wide spatial granularity of the BreBeMi area could make identifying interesting areas from permutational tests challenging. We believe that the latter hypothesis is the most likely since the analysis at the Lombardia level established a strong power of mobility weights in identifying spatial clusters.

During the second wave period, the RL projected mobility data lose explanatory power, but we have complete Trenord data about railway mobility. Considering the limitations of the area and granularity, we can interpret the absence of interesting areas together with the findings of subsection 5.2.2 at the Lombardia level. In that case, neither contiguity-based spatial weights nor mobility-based ones (based on the unreliable mobility description of RL data, which, however, could still identify global spatial autocorrelation in the epidemic phenomenon of the second wave period) could recognise interesting areas in the form of the epidemic hotspots. This fact, coupled with the exploratory analysis of the mortality rates of subsections 5.2.1 and 5.3.1, enforces the conviction that the epidemic phenomenon developed homogeneously in the area during the second wave period.

Putting together the findings revealed by local and global Moran indexes in this subection and subsection 5.3.2, no evidence of a prevalent role of public railway transport in epidemic diffusion is found in the analysis. This does not allow us to conclude that public railway transport did not play a role in this matter, but we can hypothesise that it configures itself as a part of overall mobility, following the same trends. Thus, when overall mobility shows a connection with the epidemic phenomenon, railway mobility does the same in its framework, but we have no periods of a link between mobility and epidemic found only in railway mobility.

5.4. Discussion and answer to the research questions

We now combine the points discussed in this Chapter to answer the two research questions, discussing the implications and acknowledging the limitations of our work:

- *Q1: Can we establish a link between the evolution of the pandemic during 2020 and mobility?*
- *Q2: Did railroad mobility play a relevant role in the pandemic diffusion compared to other kinds of mobility?*

To address these questions, we investigated the link between epidemic and mobility data selecting one epidemic indicator (mortality rates) and defining spatial weights from mobility data. These weights allow us to assess the relationship between their spatial description and epidemic data through global and local Moran indexes. The epidemic indicator remained fixed while we tried various mobility data to define spatial weights associated with overall or railway mobility and mobility estimated according to data prior to 2020 (projected) or relating to actual movements in 2020 (real). Figure 5.2 presents our procedure to analyse the relationship between mobility and epidemic spread, while Figure 5.3 shows the different levels of analysis we explored.

We first conducted the spatial analysis in a wide area with fine spatial granularity (Lombardia area) using projected mobility data from Regione Lombardia about overall mobility. We then restricted the same research to the BreBeMi area to compare projected and real mobility from Regione Lombardia and Trenord OD matrices. At this level, we considered the role of public railway mobility, described by the RL and Trenord data, compared with overall mobility through the RL data.

5.4.1. Q1: Can we establish a link between the evolution of the pandemic during 2020 and mobility?

Lombardia's spatial analysis shows a strong link between mobility and epidemic phenomena. The spatial mobility description provided by the RL data relates to the epidemic phenomenon in both the first and (to a lesser extent) second wave periods, identifying weeks of significantly positive spatial autocorrelation. Moreover, mobility-based spatial weights identify larger high-high areas during the first wave period than contiguity-based spatial weights. The role of the flagged high-high areas in epidemic diffusion should be investigated more deeply, establishing if these areas were starting locations of the spread through their outbound mobility flows.

From the analysis of the RL and Trenord-derived mobility data at the BreBeMi level, we have evidence confirming the existence of a link between epidemic diffusion and mobility in the first wave in the area. The connection is weaker than what we observed in the Lombardia area, likely due to the limited area and wide spatial granularity we must adopt for the analysis. Still, we can see evidence of positive spatial autocorrelation in the first wave period by all the mobility-based spatial descriptions adopted.

All the spatial analyses highlight the role of the first national lockdown in breaking the positive spatial autocorrelation. Indeed, the global Moran indexes decrease in this period. After a while, no positive spatial autocorrelation is detected in the mortality rates: the restrictions had a consequence in the spatial patterns of the mortality rates. Moreover, the high-high areas flagged in the local spatial analysis disappears sometime after the beginning of the first lockdown.

Concerning the second wave period, we should investigate the link between epidemic and mobility more deeply. Even though the RL mobility data lose their reliability in describing mobility trends after the COVID-19 outbreak, they still show some weeks of positive spatial autocorrelation with no detected hotspots at the Lombardia level. This may indicate a more homogenous diffusion trend than the first wave period, with similar mortality rates in neighbouring areas, without particular locations showing values much higher than average.

We did not find evidence supporting this interpretation in the spatial analysis of the BreBeMi area by any mobility description, particularly the one derived from actual Trenord data, the only mobility data we have representing real mobility trends for the second wave period. Indeed, we did not detect any week of positive spatial autocorrelation. Still, we have to underline that the second wave had little effect on mortality in the

BreBeMi area because of the strong impact on mortality suffered during the first wave period. The analysis of the second wave period should be repeated in a larger region, considering relevant mobility data derived from actual mobility flows.

5.4.2. Q2: Did railroad mobility play a relevant role in the pandemic diffusion compared to other kinds of mobility?

We assessed the role of railway mobility in the epidemic diffusion only in the BreBeMi area. First, we considered the RL projected data, which contains overall and railway mobility information. Then, we introduced the Trenord-derived data about actual movements by train in 8 months of 2020.

Considering the global spatial analysis, we recognised the role of railway mobility in describing the positive spatial autocorrelation in the mortality rates for a few weeks belonging to the first wave period. Thus, a connection exists between this kind of mobility and the epidemic phenomenon, as described by mortality rates.

However, no evidence was found of a prevalent role of railway mobility compared to overall mobility. To come to this conclusion, we interpreted together the RL and Trenord-derived mobility data since the Trenord analysis during the first wave period suffers from the lack of data describing January and February's mobility. Without dynamic OD matrices describing these months, we supplemented the railway description with the static projected RL data. Indeed, we have no reason to believe that mobility in the first months of 2020 should differ from the RL data. Thus, we have a complete representation of railway mobility to draw conclusions. The RL projected data individuates only one significant week for railway mobility, which was one of the two weeks identified by overall mobility-based spatial weights. The Trenord data allow us to highlight another week of positive spatial autocorrelation according to railway mobility. Moreover, the global Moran indexes induced by railway mobility (according to the two mobility descriptions) are always lower than the ones computed considering overall mobility through the RL data.

Considering the second wave period, we observe no positive spatial autocorrelation by either mobility description. Remarkably, the Trenord data provided us with the only reliable mobility representation for this period since the RL data provides a static projected mobility description. Thus, in the BreBeMi area (which was lightly affected in the second wave period), there is no evidence of positive spatial autocorrelation in the epidemic phenomenon described by railway mobility. Still, the limited extension and wide granularity of the area analysed could make identifying spatial autocorrelation challenging. The robustness of these findings should be confirmed considering a larger region.

Railway mobility-based spatial weights identify no interesting areas in the BreBeMi area. However, overall mobility still fails in detecting significant local Moran indexes, while the same weights identified extensive high-high zones at the Lombardia level. Indeed, the limited extension and wide spatial granularity probably caused this failure in the BreBeMi area. Thus, we should investigate deeper the role of railway mobility compared to overall mobility in detecting interesting areas.

Conclusion

This work was motivated by the interest in analysing mobility data to assess the role of mobility in COVID-19 spread through 2020, focusing on evaluating if a particular kind of public transport (public railway transport) had a prevalent role in the epidemic's diffusion.

Achievements

To reach our research goal, we collected a mobility dataset (the RL OD matrix) describing projected movements between municipalities in Lombardia, estimated from data collected prior to 2020. This dataset was the starting point of our analysis. Still, we needed more data dynamically describing mobility flows during 2020 to consider the impact that the disease spread and restrictions adopted to contain it had on mobility, particularly on public transport. We derived a representation of public railway mobility from data provided by Trenord, the local railway company.

The need for mobility data accurately describing mobility flows in 2020 led us to develop a procedure to estimate weekly OD matrices describing movements by train in the limited portion of the Trenord network covered by the available data. Our estimation pipeline presents some novelties compared to other works in trip distribution modelling to deal with problems specific to the Trenord datasets, like estimating trips from tickets and correcting passenger counts for train rides with missing data. The estimated Trenord OD matrices obtained by applying the pipeline showed coherency with some reality-induced principles and gave us qualitative evidence of their ability to represent actual mobility flows in 2020. The estimation pipeline's relevance goes beyond the scope of our research, as it could be generalised and extended to the whole Trenord network or other transportation networks, provided that we could derive seed OD matrices and estimates of trips beginning and ending in each zone of the network. Comparison with the RL mobility data showed a strong predictive power of the railway RL data in actual Trenord-derived data.

Regarding our research goal, the purpose of the Trenord-derived mobility data in this work is twofold: they describe actual mobility trends through 2020 and allow us to address the role of a specific kind of public transport believed to be a primary carrier in epidemic diffusion.

Coming to the analysis of the relationship between mobility and epidemics, we developed a procedure to analyse spatial autocorrelation in an epidemic feature, mortality rates in our case, through a spatial description based on mobility flows. To apply this procedure, we need an areal indicator describing the epidemic outcomes and mobility data in the form of OD matrices. In our work, we chose mortality rates as the epidemic indicator and computed mobility-based spatial descriptions through the static RL OD matrix or the dynamic Trenord ones. We considered two regions, first the entire Lombardia region and then a limited area covered by the Trenord data, named BreBeMi area.

Results indicate a strong relationship between mobility and epidemic spread in Lombardia through 2020. All mobility data considered always revealed periods of positive spatial autocorrelation in the mortality rates. Moreover, the mobility-based spatial description identified larger high-high spatial clusters in the Lombardia area during the first wave period than the spatial description based on zones' contiguity. These areas warrant further investigation into their potential role in the epidemic's spread. This first level of analysis highlighted the link of mobility-based spatial weights to the epidemic phenomenon and called for more insight into mobility trends.

Concerning the role of public railway transport, we investigated this aspect only in the BreBeMi area to exploit the estimated dynamic Trenord data. We had to supplement the description given by the Trenord data with the static RL data describing railway mobility because we missed Trenord data for the months of January and February, which we believe had the highest impact on the disease diffusion following the COVID-19 outbreak in the area. We observed that railway mobility-based spatial weights could still identify some weeks of positive spatial autocorrelation in the epidemic feature. On the other hand, the identified period matched the period highlighted by the overall mobility description and we never found positive spatial autocorrelation according to the railway mobility-based spatial description in periods of no correlation with overall mobility. Moreover, the global Moran indexes computed according to the railway mobility spatial description were consistently lower than the ones computed according to overall mobility.

This thesis has shown the potential power spatial descriptions based on mobility could have in identifying epidemic diffusion's periods and hotspots. Concerning the role of public transport, this work found no evidence to affirm that a particular kind of public transport (railway mobility) played a prevalent role in the pandemic's diffusion compared to overall mobility. Further research is needed to assess if mobility-based spatial descriptions could describe the epidemic diffusion's dynamic considering different kinds of mobility. However, this analysis might be the starting point for developing decision tools for policymakers having to take action to react to epidemic phenomena, supported by further research about the role of mobility in spreading the disease. Moreover, if we could affirm that public transport had no prevalent contribution to the epidemic's spread, restrictions on its usage for future outbreaks could be reviewed. This could release businesses in the field of huge losses and encourage people not to fear this transportation method.

Future developments

While this work has made significant contributions to understanding the role of mobility in epidemic diffusion, we acknowledge limitations to our analyses, paving the way for future research. This Section will outline some potential future developments for our work.

First, even if past works show the representative power of the mortality rate of people aged 70 or more to describe the pandemic evolution during 2020, the spatial analysis pipeline we developed is general. Thus, we could repeat the study considering different kinds of areal epidemic data, like the rate of positive swabs, quarantined people or occupancy of intensive care units to represent various aspects of the epidemic. Another point to consider if we want to analyse the pandemic evolution through 2021 is that after the end of 2020, epidemic waves show their effect on peak infection rates but hardly on mortality rates, as shown in [11]. Because of this, it could make sense to repeat the spatial analysis with a different epidemic indicator than the mortality rate when considering later epidemic periods.

At the same time, we could also consider different mobility data in the spatial analysis pipeline. Concerning this point, we underline that the estimation procedure we developed to estimate movements by train in the Trenord network could be generalised to other transportation networks, such as metro or bus systems or other train networks. Alternatively, we could extend the area of estimation to include the whole Trenord network instead of a limited portion and repeat the spatial analysis with little effort. Another future development we cited multiple times is the extension of the estimation of trips by train in the BreBeMi area to include January and February railway mobility. This extension could allow us to analyse actual movements by train in the spatial analysis thoroughly. However, we expect to obtain mobility trends similar to those described by the projected RL data, as we have no reason to believe that actual mobility should differ from the projected description in this period.

In Chapter 5, we compared mobility-based spatial weights with contiguity-based ones in the Lombardia area. We could extend the comparison to include distance-based weights, like inverse distance or distance bands. These weights could also add a layer to the analysis of the BreBeMi area, for which we could not define contiguity-based spatial weights due to the lack of complete representation of neighbours for the zones of the study.

Another strand of research could be assessing the mobility trends' changes in Lombardia and comparing them with the projected mobility description for 2020 released by RL Open Data in 2019. We stated multiple times that we believe mobility has completely changed since the epidemic. However, the comparison between railway mobility (as described by the projected RL data and Trenord actual data) points to the predictive power of the projected data for the Trenord-derived mobility trends. We investigated this matter only for railway mobility in the BreBeMi area, but we could repeat and extend the comparative analysis if we estimated other kinds of mobility data or the same data in other areas and periods. Moreover, the mobility-based spatial weights derived from the projected RL data could still identify positive spatial autocorrelation in the epidemic feature during the second wave period, which hints again at the retention of validity of these data to provide insights on mobility trends.

Finally, in Chapter 5, we declared that the larger high-high spatial clusters identified by mobility weights called for further investigation into their role in epidemic diffusion. We could elaborate on this matter by analysing the time series of the epidemic feature for two nearby areas, according to the notion of nearness induced by mobility. In particular, we would choose a high-high area (identified in the local spatial analysis) and another nearby zone (looking for high values in the mobility-based spatial weights) to identify anticipators in the first time series for the second one, applying time-series methods like the Engle-Granger causality test [83]. Suppose we established that the first area influenced the second one. In that case, we could go on to investigate their relationship through the first outbreak and lockdown period to assess the role of lockdown on their connection.

Bibliography

[1] Elena Dusi. Misterioso virus in Cina, 59 colpiti da polmonite. *Repubblica*, January 2020. URL https://www.repubblica.it/salute/2020/01/06/news/misterioso_virus_in_cina_59_colpiti_da_polmonite-245096236/. Last accessed 2023-04-13.

[2] Alessandra Corica, Oriana Liso and Mauro Rancati. Coronavirus, i contagi nel Lodigiano sono 15: i primi sono un 38enne di Codogno e sua moglie. In isolamento 250 persone. *Repubblica*, February 2020. URL https://milano.repubblica.it/cronaca/2020/02/21/news/coronavirus_a_milano_contaggiato_38enne_e_un_italiano_ricoverato_a_codogno-249121707/. Last accessed 2023-04-13.

[3] Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao et al. Coronavirus pandemic (COVID-19). *Our World in Data*, 2020. URL https://ourworldindata.org/coronavirus. Last accessed 2023-04-13.

[4] Istituto Superiore di Formazione e Ricerca per i Trasporti. L'impatto del lockdown sui comportamenti di mobilità degli italiani, May 2020. URL https://www.isfort.it/progetti/limpatto-del-lockdown-sui-comportamenti-di-mobilita-degli-italiani/. Last accessed 2023-04-13.

[5] Marco Morino. Effetto long COVID sul trasporto pubblico locale: nel 2022 passeggeri giù del 21%. *Il Sole 24 Ore*, February 2022. URL https://www.ilsole24ore.com/art/effetto-long-covid-tpl-2022-passeggeri-giu-21percento-AEgp5WdC. Last accessed 2023-04-13.

[6] Filipa Sà. Socioeconomic determinants of COVID-19 infections and mortality: evidence from England and Wales. *IZA Policy Paper*, 159, May 2020.

[7] Mattia Borsati, Silvio Nocera and Marco Percoco. Questioning the spatial association between the spread of COVID-19 and transit usage in Italy. Technical Report 11, eng. Working Paper Series, 2020.

[8] Claudio Riccardi. I benefici del trasporto pubblico. *Tutto Green*, 2013. URL https:

//www.tuttogreen.it/i-benefici-del-trasporto-pubblico/. Last accessed 2023-04-13.

[9] ISTAT. Decessi e cause di morte: cosa produce l'ISTAT, 2022. Data retrieved from ISTAT, URL https://www.istat.it/it/archivio/240401. Last accessed 2023-04-13.

[10] Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli and Piercesare Secchi. A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. *Spatial Statistics*, 49:100541, 2022.

[11] Veronica Mazzola. The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021. Master's thesis, Politecnico di Milano, 2022.

[12] Stefano Maria Iacus, Carlos Santamaria, Francesco Sermi et al. Human mobility and COVID-19 initial dynamics. *Nonlinear Dynamics*, 101(3):1901–1919, 2020.

[13] Regione Lombardia Open Data. Il programma Open Data, 2023. https://hub.dati.lombardia.it/stories/s/q7aq-wkga. Last accessed 2023-04-13.

[14] Regione Lombardia Open Data. Storia di Open Data, 2023. https://dati.lombardia.it/stories/s/Storia-di-Open-Data/n767-ptkf. Last accessed 2023-04-13.

[15] Trenord. Chi siamo, 2023. URL https://www.trenord.it/chi-siamo/. Last accessed 2023-04-13.

[16] Lab24. Cose che noi umani, 2021. URL https://lab24.ilsole24ore.com/storia-coronavirus/. Last accessed 2023-04-13.

[17] Silvia Camporesi, Federica Angeli and Giorgia Dal Fabbro. Mobilization of expert knowledge and advice for the management of the Covid-19 emergency in Italy in 2020. *Humanities and Social Sciences Communications*, 9:54, February 2022.

[18] Protezione Civile. Dati COVID-19 Italia, 2023. Data retrieved from Protezione Civile, URL https://github.com/pcm-dpc/COVID-19. Last accessed 2023-04-13.

[19] Ministero della Salute. COVID-19 - Situazione nel mondo, 2023. URL https://www.salute.gov.it/portale/nuovocoronavirus/dettaglioContenutiNuovoCoronavirus.jsp?area=nuovoCoronavirus&id=5338&lingua=italiano&menu=vuoto. Last accessed 2023-04-13.

[20] Grazia Labate. Cosa è cambiato dopo il Covid? *quotidianosanità.it*, November 2022.

URL https://www.quotidianosanita.it/studi-e-analisi/articolo.php?articolo_id=108669. Last accessed 2023-04-13.

[21] Redazione online. Coronavirus, in isolamento Castiglione d'Adda, Codogno, Casalpusterlengo e altri 7 comuni. *Corriere della Sera*, February 2020. URL https://milano.corriere.it/notizie/cronaca/20_febbraio_21/coronavirus-castiglione-d-adda-codogno-isolamento-non-uscite-casa-non-andate-pronto-soccorso-4b0597ee-548e-11ea-9196-da7d305401b7.shtml. Last accessed 2023-04-13.

[22] Redazione online. Coronavirus a Bergamo, i casi raddoppiano in un giorno: sono più di 200. Positivo anche il sindaco di Nembro Claudio Cancelli. *Corriere della Sera*, March 2020. URL https://bergamo.corriere.it/notizie/cronaca/20_marzo_01/coronavirus-bergamo-positivo-anche-sindaco-nembro-claudio-cancelli-salta-bergamo-film-meeting-4b8464ce-5bc9-11ea-ae74-e93752023e91.shtml. Last accessed 2023-04-13.

[23] Adriana Logroscino. Inchiesta Covid, Alzano Lombardo e Nembro: discussioni e contrasti. Il Cts propose la zona rossa all'unanimità. *Corriere della Sera*, March 2023. URL https://www.corriere.it/politica/23_marzo_04/inchiesta-covid-0f219dbe-b9f7-11ed-95ad-287c8c9a1773.shtml. Last accessed 2023-04-13.

[24] Alessandro Gatta. Coronavirus, primo caso nel Bresciano già a gennaio: dove tutto è cominciato. *Brescia Today*, March 2020. URL https://www.bresciatoday.it/social/coronavirus-dove.html. Last accessed 2023-04-13.

[25] R. A. Treni Trenord in Lombardia: tutte le variazioni e cancellazioni fino all'8 marzo 2020. *Milano Today*, March 2020. URL https://www.milanotoday.it/attualita/variazioni-treni-trenord-8-marzo-2020.html. Last accessed 2023-04-13.

[26] Marco Morino. Ferrovie, l'emergenza Covid taglia il 98% dei treni veloci. *Il Sole 24 Ore*, March 2020. URL https://www.ilsole24ore.com/art/ferrovie-l-emergenza-covid-taglia-98percento-treni-veloci-ADPlSIF. Last accessed 2023-04-13.

[27] Jon Henley. Italy records lowest coronavirus death toll for a week. *The Guardian*, April 2020. URL https://www.theguardian.com/world/2020/apr/01/italy-extends-lockdown-amid-signs-coronavirus-infection-rate-is-easing. Last accessed 2023-04-13.

[28] Redazione. Coronavirus, Trenord e Atm da lunedì 4 maggio: la Fase 2 dei trasporti in Lombardia. *Il Cittadino*, May 2020. URL https://www.ilcittadinomb.it/new

s/cronaca/coronavirus-trenord-e-atm-da-lunedi-4-maggio-la-fase-2-dei-trasporti-in-lombardia/. Last accessed 2023-04-13.

[29] Redazione. Coronavirus, Fase 2: ecco come sarà viaggiare in treno. *Brescia Today*, May 2020. URL https://www.bresciatoday.it/motori/mobilita-sostenibile/coronavirus-fase-2-treno.html. Last accessed 2023-04-13.

[30] Redazione. Treni, dal 14 tutti i posti si possono occupare, "Massima offerta nelle ore di punta". *L'Eco di Bergamo*, September 2020. URL https://www.ecodibergamo.it/stories/bergamo-citta/treni-dal-14-tutti-i-posti-si-possono-occuparemassima-offerta-nelle-ore-di-pun_1370488_11/. Last accessed 2023-04-13.

[31] Matteo Corner. Il problema dei mezzi pubblici in tempi di pandemia. *Il Post*, October 2020. URL https://www.ilpost.it/2020/10/19/mezzi-pubblici-coronavirus/. Last accessed 2023-04-13.

[32] Fiorenza Sarzanini. Nuovo Dpcm di ottobre: mascherine all'aperto, divieto di ballo e limiti alle feste. *Corriere della Sera*, October 2020. URL https://www.corriere.it/cronache/20_ottobre_06/nuovo-dpcm-ottobre-2020-covid-c6d6f5ec-071d-11eb-a92a-d6e5260ddebb.shtml. Last accessed 2023-04-13.

[33] Monica Guerzoni and Fiorenza Sarzanini. Nuovo Dpcm, coprifuoco alle 22, nelle zone rosse lockdown di 15 giorni: il testo valido fino al 3 dicembre. *Corriere della Sera*, November 2020. URL https://www.corriere.it/cronache/20_novembre_03/nuovo-dpcm-oggi-coprifuoco-32c653a6-1dc1-11eb-9970-42ca5768e0fd.shtml. Last accessed 2023-04-13.

[34] Andrea Gagliardi. Regioni, ecco chi passa in zona arancione e chi in zona gialla. *Il Sole 24 Ore*, November 2020. URL https://www.ilsole24ore.com/art/regioni-lombardia-piemonte-ecco-chi-passa-zona-arancione-e-chi-rischia-rossa-AD2Qdi4?refresh_ce=1. Last accessed 2023-04-13.

[35] Redazione. Coronavirus: regioni rosse, arancioni e gialle. Cosa è vietato in ogni fascia, scarica l'autocertificazione. *Repubblica*, November 2020. URL https://www.repubblica.it/politica/2020/11/04/news/scheda_dpcm_covid-273036534/. Last accessed 2023-04-13.

[36] Songhua Hu, Chenfeng Xiong, Mofeng Yang et al. A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic. *Transportation Research Part C: Emerging Technologies*, 124:102955, 2021.

[37] Samuel Engle, John Stromme and Anson Zhou. Staying at Home: Mobility Effects of COVID-19. *SSRN*, 2020.

[38] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli et al. Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, 117(27):15530–15535, 2020.

[39] European Commission, Joint Research Centre, Michele Vespe et al. *Mobility and economic impact of COVID-19 restrictions in Italy using mobile network operator data*. Publications Office, 2021.

[40] Carlos Santamaria, Francesco Sermi, Spyridon Spyratos et al. Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis. *Safety Science*, 132:104925, 2020.

[41] Hamada S. Badr, Hongru Du, Maximilian Marshall et al. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.

[42] Nishant Kishore, Rebecca Kahn, Pamela P. Martinez et al. Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of SARS-CoV-2. *Scientific Reports*, 11(1):6995, 2021.

[43] Valentina Pieroni, Angelo Facchini and Massimo Riccaboni. COVID-19 vaccination and unemployment risk: lessons from the Italian crisis. *Scientific Reports*, 11(1):18538, 2021.

[44] Mustafa Tevfik Kartal, Özer Depren and Serpil Kiliç Depren. The relationship between mobility and COVID-19 pandemic: Daily evidence from an emerging country by causality analysis. *Transportation Research Interdisciplinary Perspectives*, 10:100366, 2021.

[45] Matilde Bonadia. Reshaping a nation: mobility, commuting and contact patterns during the COVID-19 outbreak. The Trenord case. Master's thesis, Politecnico di Milano, 2021.

[46] Nuriye Melisa Bilgin. Tracking COVID-19 spread in Italy with Mobility Data. Technical report, Koc University-TUSIAD Economic Research Forum, 2020.

[47] Astrid Krenz and Holger Strulik. The benefits of remoteness - digital mobility data, regional road infrastructure, and COVID-19 infections. *German Economic Review*, 22(3):257–287, 2021.

[48] Behzad Vahedi, Morteza Karimzadeh and Hamidreza Zoraghein. Spatiotemporal prediction of COVID-19 cases using inter- and intra-county proxies of human interactions. *Nature Communications*, 12(1):6440, 2021.

[49] Cornelia Ilin, Sébastien Annan-Phan, Xiao Hui Tai et al. Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific Reports*, 11(1):13531, 2021.

[50] Pengyu Zhu, Jiarong Li and Yuting Hou. Applying a Population Flow-Based Spatial Weight Matrix in Spatial Econometric Models: Conceptual Framework and Application to COVID-19 Transmission Analysis. *Annals of the American Association of Geographers*, 112(8):2266–2286, 2022.

[51] Mattia Borsati, Michele Cascarano and Marco Percoco. Resilience to health shocks and the spatial extent of local labour markets: evidence from the Covid-19 outbreak in Italy. *Regional Studies*, pages 1–18, 2022.

[52] Francesco Checchi and Les Roberts. Interpreting and using mortality data in humanitarian emergencies. Technical report, 2005. `https://odihpn.org/publication/interpreting-and-using-mortality-data-in-humanitarian-emergencies/`. Last accessed 2023-04-13.

[53] Matteo Chinazzi, Jessica T. Davis, Marco Ajelli et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489):395–400, 2020.

[54] Ritabrata Dutta, Susana N. Gomes, Dante Kalise and Lorenzo Pacchiardi. Using mobility data in the design of optimal lockdown strategies for the COVID-19 pandemic. *PLOS Computational Biology*, 17:1–25, 2021.

[55] Wen-Hao Chiang, Xueying Liu and George Mohler. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting*, 38(2):505–520, 2022.

[56] Benjamin Lucas, Behzad Vahedi and Morteza Karimzadeh. A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. *International Journal of Data Science and Analytics*, 2022.

[57] Regione Lombardia. Matrice OD2020 - passeggeri, 2019. Data retrieved from Regione Lombardia Open Data, URL `https://www.dati.lombardia.it/Mobilit-e-trasporti/Matrice-OD2020-Passeggeri/hyqr-mpe2`. Last accessed 2023-04-13.

[58] Regione Lombardia. Matrice Origine/Destinazione: i dati sulle abitudini di sposta-

mento in Lombardia, 2015. URL https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioServizio/servizi-e-informazioni/Imprese/Imprese-di-trasporto-e-logistica/ser-matrice-od-infr/matrice-od. Last accessed 2023-04-13.

[59] ISTAT. Popolazione e famiglie, 2020. Data retrieved from ISTAT, URL https://www.istat.it/it/popolazione-e-famiglie?dati. Last accessed 2023-04-13.

[60] Trenord. Linee regionali, 2022. URL https://www.trenord.it/linee-e-orari/il-nostro-servizio/linee-regionali/. Last accessed 2023-04-13.

[61] Trenord. Linee e orari, 2022. https://www.trenord.it/linee-e-orari/.

[62] Trenord. Il bilancio di sostenibilità 2021, 2022. https://www.trenord.it/chi-siamo/bds/.

[63] Ivano Pinna and Bruno Dalla Chiara. Automatic passenger counting and vehicle load monitoring. *Ingegneria Ferroviaria*, 65:101–138, 2010.

[64] ISTAT. Matrici di contiguità, distanza e pendolarismo, 2019. Data retrieved from ISTAT, URL https://www.istat.it/it/archivio/157423. Last accessed 2023-04-13.

[65] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022.

[66] Johan Barthélemy and Thomas Suesse. mipfp: An R Package for Multidimensional Array Fitting and Simulating Multivariate Bernoulli Distributions. *Journal of Statistical Software, Code Snippets*, 86(2):1–20, 2018.

[67] Roger Bivand and David W.S. Wong. Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3):716–748, 2018.

[68] Andrew W. Evans. Some properties of trip distribution methods. *Transportation Research*, 4:19–36, 1970.

[69] S. M. Macgill. Theoretical Properties of Biproportional Matrix Adjustments. *Environment and Planning A: Economy and Space*, 9(6):687–701, 1977.

[70] Agostino Torti, Marta Galvani, Valeria Urbano et al. Analysing transportation system reliability: the case study of the metro system of Milan. Technical Report 84, MOX-Report No. 84/2021, 2021. URL https://www.mate.polimi.it/biblioteca/add/qmox/84-2021.pdf. Last accessed 2023-04-13.

[71] Mark R. McCord, Rabi G. Mishalani, Prem Goel and Brandon Strohl. Iterative Proportional Fitting Procedure to Determine Bus Route Passenger Origin–Destination Flows. *Transportation Research Record*, 2145(1):59–65, 2010.

[72] Joseph B. Lang. Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics*, 32(1):340 – 383, 2004.

[73] Abdoul-Ahad Choupani and Amir Reza Mamdoohi. Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation research procedia*, 17:223–233, 2016.

[74] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[75] Trenord. Biglietti e abbonamenti, 2022. URL `https://www.trenord.it/biglietti/`. Last accessed 2023-04-13.

[76] Google Maps. Driving directions from Monte Isola to Sulzano, 2021. Data retrieved from Google Maps, URL `https://www.google.com/maps`. Last accessed 2023-04-13.

[77] Patrick A. P. Moran. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, 10(2):243–251, 1948.

[78] Luc Anselin. Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2):93–115, 1995.

[79] Roger S. Bivand, Edzer Pebesma and Virgilio Gòmez-Rubio. *Applied Spatial Data Analysis with R*. Springer, 2013.

[80] Luc Anselin, Ibnu Syabri and Youngihn Kho. Geoda: An introduction to spatial data analysis. *Geographical Analysis*, 38(1):5–22, 2006.

[81] R. P. Haining. Spatial autocorrelation. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 14763–14768. Pergamon, Oxford, 2001.

[82] Luc Anselin. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, pages 111–125. Taylor & Francis, London, 1996.

[83] Clive W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969.

A | Appendix A

This Appendix aims to explain in detail our reasoning in selecting the time aggregation needed to define the epidemic indicator $m_i^{[w]}$ on which we based the spatial analysis of Chapter 5.

First, we recall the formula of the epidemic indicator used to model the epidemic response, expressed in Equation (5.1):

$$m_i^{[w]} = \frac{\sum_{q=w-slide+1}^{q=w} deaths_i^{[q]}}{population_i}$$

In the spatial analysis pipeline presented in Chapter 5 (see Figure 5.2 for a brief reminder of the procedure), we compute global and local Moran indexes to assess global and local spatial autocorrelation in the epidemic feature. We explore four values of the *slide* parameter which express the time aggregation of the mortality index, $slide \in [1, 2, 3, 4]$ and motivate our reasons for selecting $slide = 2$. We conduct this analysis in the Lombardia area in the same fashion of subsection 5.2.2. We consider mobility-based spatial weights derived from the RL OD matrix.

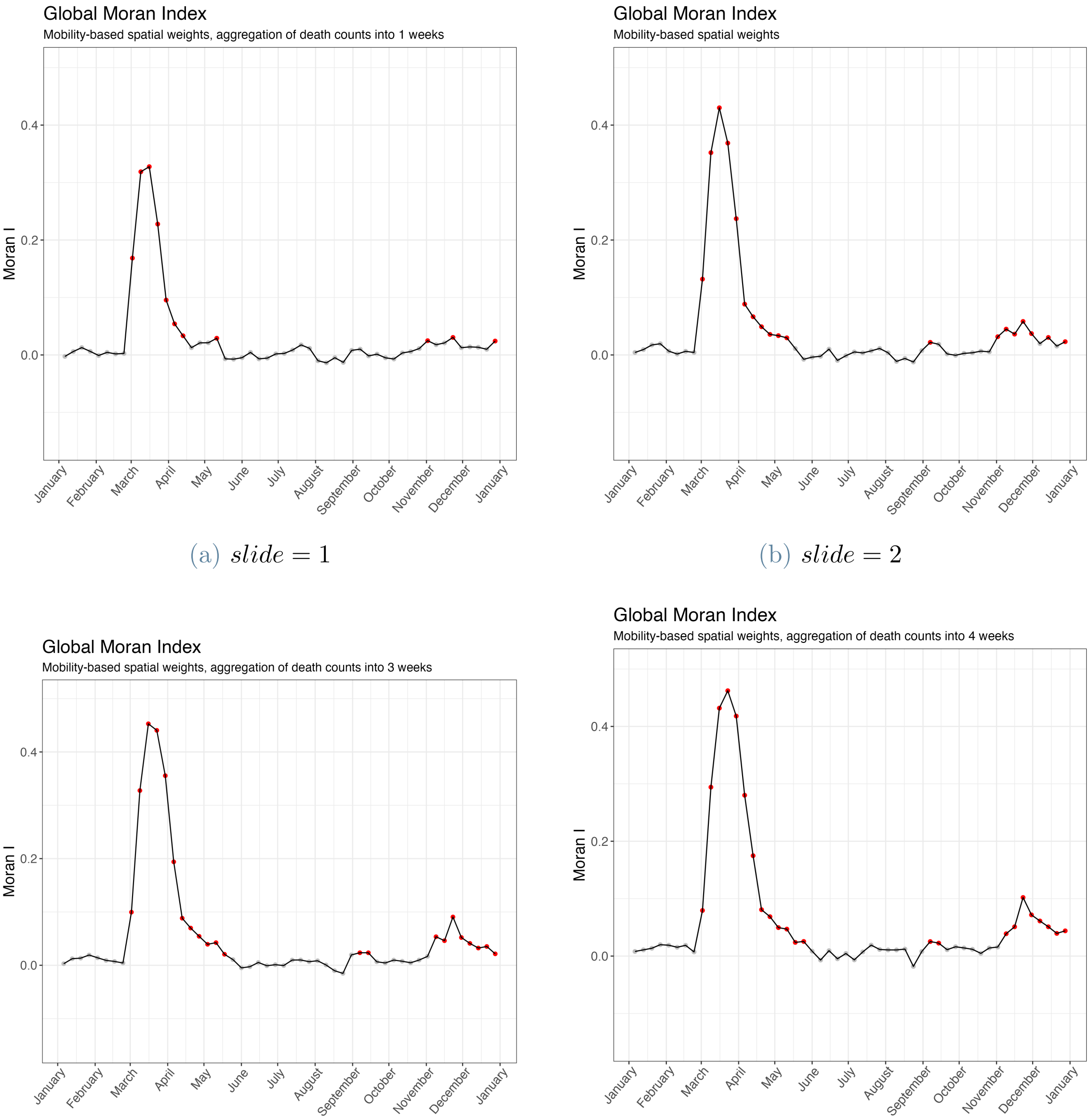
The rationales which will guide us in the selection are the following:

- We want to select a *slide* large enough to smooth out weekly oscillations and highlight longer-term trends in mortality and Moran indexes.
- We want to select a *slide* small enough to identify the earliest possible variations in the mortality rates and Moran indexes.

We have to consider both the values of global and local Moran indexes in the selection.

A.1. Selection of the time aggregation parameter

Figure A.1 compares the global Moran indexes computed for every week of 2020, selecting $slide \in [1, 2, 3, 4]$.



(a) *slide* = 1

(b) *slide* = 2

(c) *slide* = 3

(d) *slide* = 4

Figure A.1: Global Moran indexes at Lombardia level based on RL mobility data, comparing $slide \in [1, 2, 3, 4]$

We can immediately discard the value $slide = 1$ because of the lower values of the indexes compared to the other cases and because of its limited power in revealing positive spatial autocorrelation in the second wave period. On the other hand, global Moran indexes give us no strong reason for selecting one of the other three possible *slide* values.

In the next pages, we display the LISA maps showing the interesting areas detected by local Moran indexes, comparing $slide \in [1, 2, 3, 4]$. We show the maps for the weeks corresponding to the first pandemic outbreak and the following lockdown period, $w \in$

$[8, 9, 10, 11, 12, 13]$. The rows of the figures represent the same week on the same page, while the columns compare different values of *slide*. In particular, Figure A.2 shows $I_i^{[w]}$ for $w \in [8, 9, 10]$ and $slide \in [1, 2]$, Figure A.3 shows $I_i^{[w]}$ for $w \in [8, 9, 10]$ and $slide \in [3, 4]$, Figure A.4 shows $I_i^{[w]}$ for $w \in [11, 12, 13]$ and $slide \in [1, 2]$, and lastly, Figure A.5 shows $I_i^{[w]}$ for $w \in [11, 12, 13]$ and $slide \in [3, 4]$.

We make some comments which guide us to the choice of the final *slide* value:

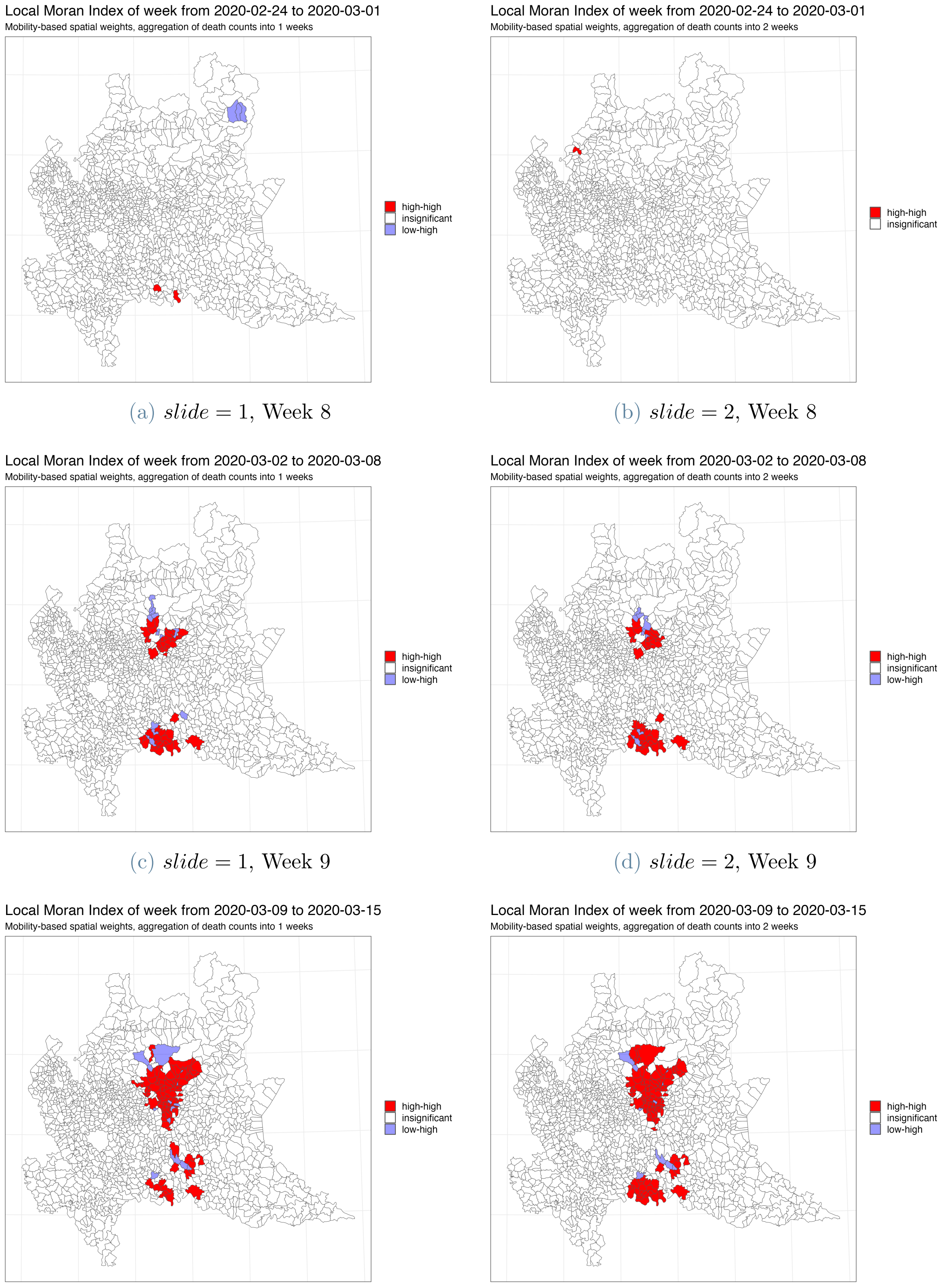
- Indexes computed for $slide = 1$ reveal a much lower number of high-high areas than the other cases. Thus, we judge that considering only one week in the computation of the mortality index does not lead to an accurate representation of the clustered spatial pattern. However, the $slide = 1$ value has already been judged unsatisfactory in the analysis of the global Moran indexes. The only point in its favour is the fact that it is the only index flagging two high-high areas in week 8, corresponding to two territories in the immediate surroundings of Codogno.

- The $slide = 2$ case highlights the highest number of high-high areas in weeks 9 and 10. Starting from week 11, the areas corresponding to the surrounding of Codogno disappear from the map while they remain significant for the $slide = 3$ and $slide = 4$ cases.

- Cases $slide = 3$ and $slide = 4$ are extremely similar. The only difference between them is the lower value of high-high areas for the $slide = 3$ case in weeks 12 and 13, compared to $slide = 4$.

Thus, we select $slide = 2$ for its power in detecting high-high areas at the earliest. The fact that the Codogno area disappears from the maps earlier than the others areas is coherent with the early restrictions adopted there, which caused reductions in mortality rates before other zones.

However, a point in favour of the $slide = 3$ (or $slide = 4$) case is the higher global Moran indexes detected during the second wave period compared to the $slide = 2$ value. Because of this, we decided to make more precise comments on the comparison between global Moran indexes computed according to mobility-based spatial weights or contiguity-based ones in the next Section of this Appendix.


Local Moran Index of week from 2020-02-24 to 2020-03-01
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-02-24 to 2020-03-01
Mobility-based spatial weights, aggregation of death counts into 2 weeks
🟥 high-high
□ insignificant
Local Moran Index of week from 2020-03-02 to 2020-03-08
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-02 to 2020-03-08
Mobility-based spatial weights, aggregation of death counts into 2 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Mobility-based spatial weights, aggregation of death counts into 2 weeks
🟥 high-high
□ insignificant
🟦 low-high

(a) $slide = 1$, Week 8

(b) $slide = 2$, Week 8

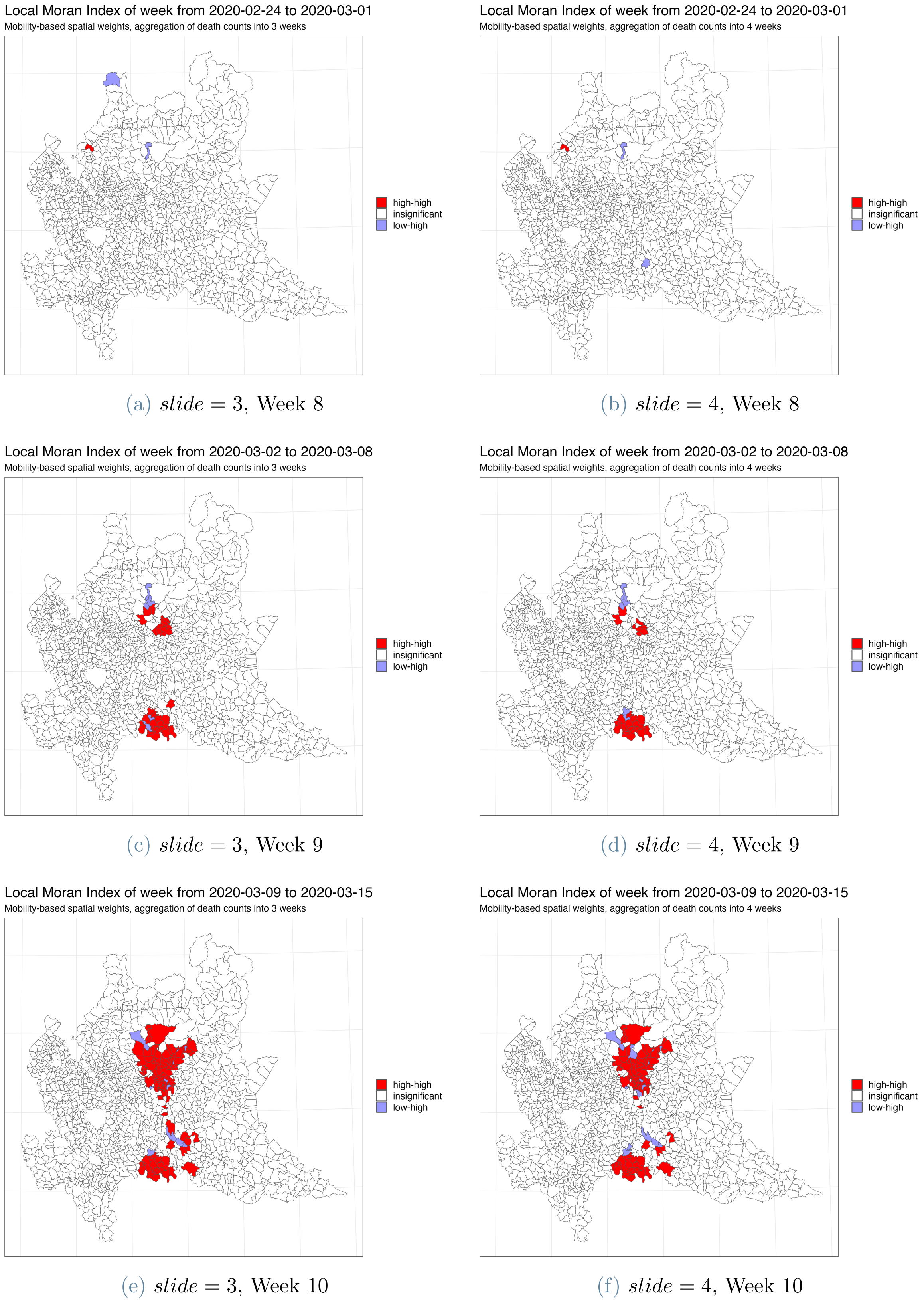
(c) $slide = 1$, Week 9

(d) $slide = 2$, Week 9

(e) $slide = 1$, Week 10

(f) $slide = 2$, Week 10

Figure A.2: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia, comparing $slide \in [1, 2]$ and $w \in [8, 9, 10]$


Local Moran Index of week from 2020-02-24 to 2020-03-01
Mobility-based spatial weights, aggregation of death counts into 3 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-02-24 to 2020-03-01
Mobility-based spatial weights, aggregation of death counts into 4 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-02 to 2020-03-08
Mobility-based spatial weights, aggregation of death counts into 3 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-02 to 2020-03-08
Mobility-based spatial weights, aggregation of death counts into 4 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Mobility-based spatial weights, aggregation of death counts into 3 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-09 to 2020-03-15
Mobility-based spatial weights, aggregation of death counts into 4 weeks
🟥 high-high
□ insignificant
🟦 low-high

(a) $slide = 3$, Week 8

(b) $slide = 4$, Week 8

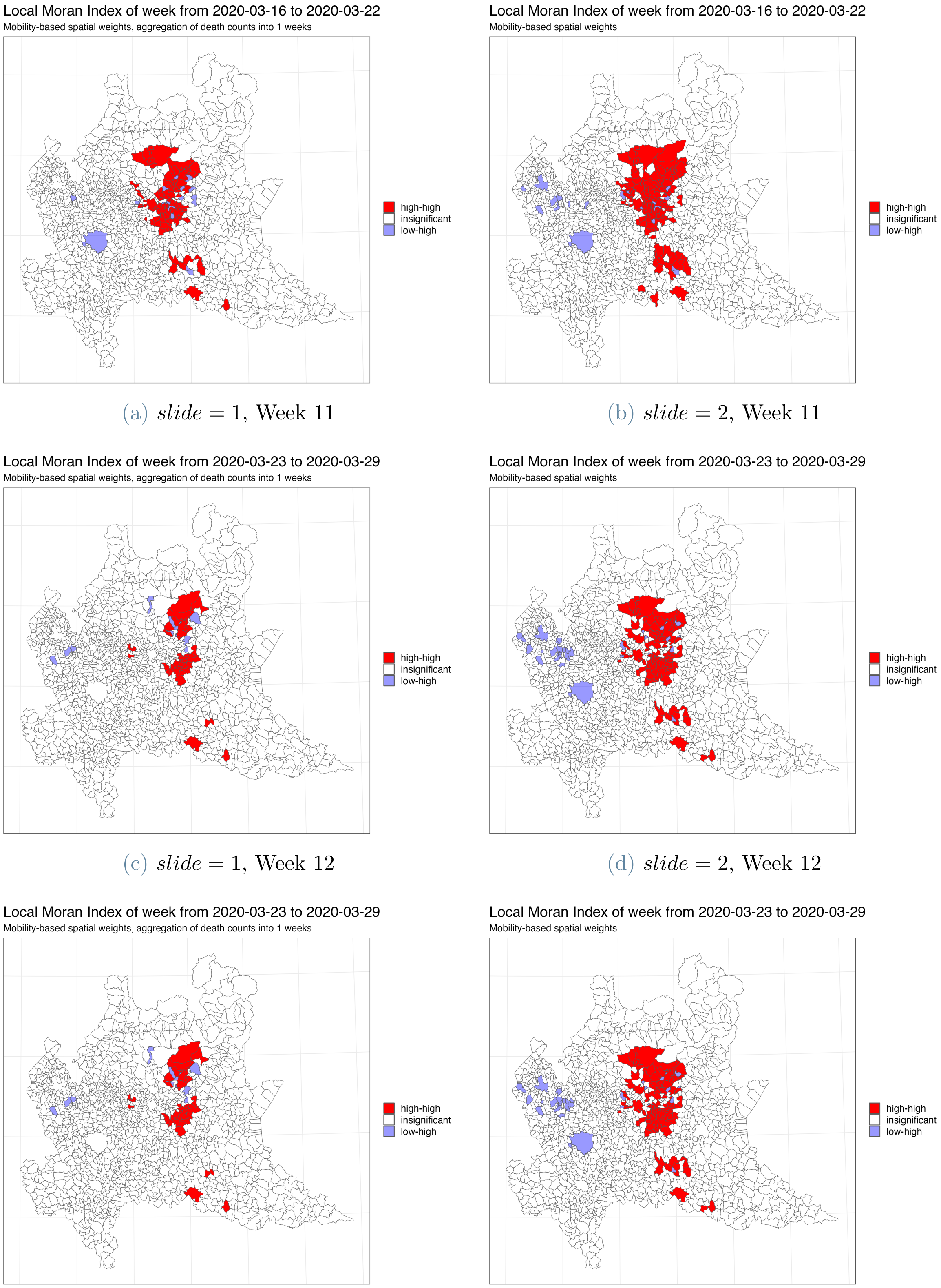
(c) $slide = 3$, Week 9

(d) $slide = 4$, Week 9

(e) $slide = 3$, Week 10

(f) $slide = 4$, Week 10

Figure A.3: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia, comparing $slide \in [3, 4]$ and $w \in [8, 9, 10]$


Local Moran Index of week from 2020-03-16 to 2020-03-22
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-16 to 2020-03-22
Mobility-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 1 weeks
🟥 high-high
□ insignificant
🟦 low-high
Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights
🟥 high-high
□ insignificant
🟦 low-high

(a) $slide = 1$, Week 11

(b) $slide = 2$, Week 11

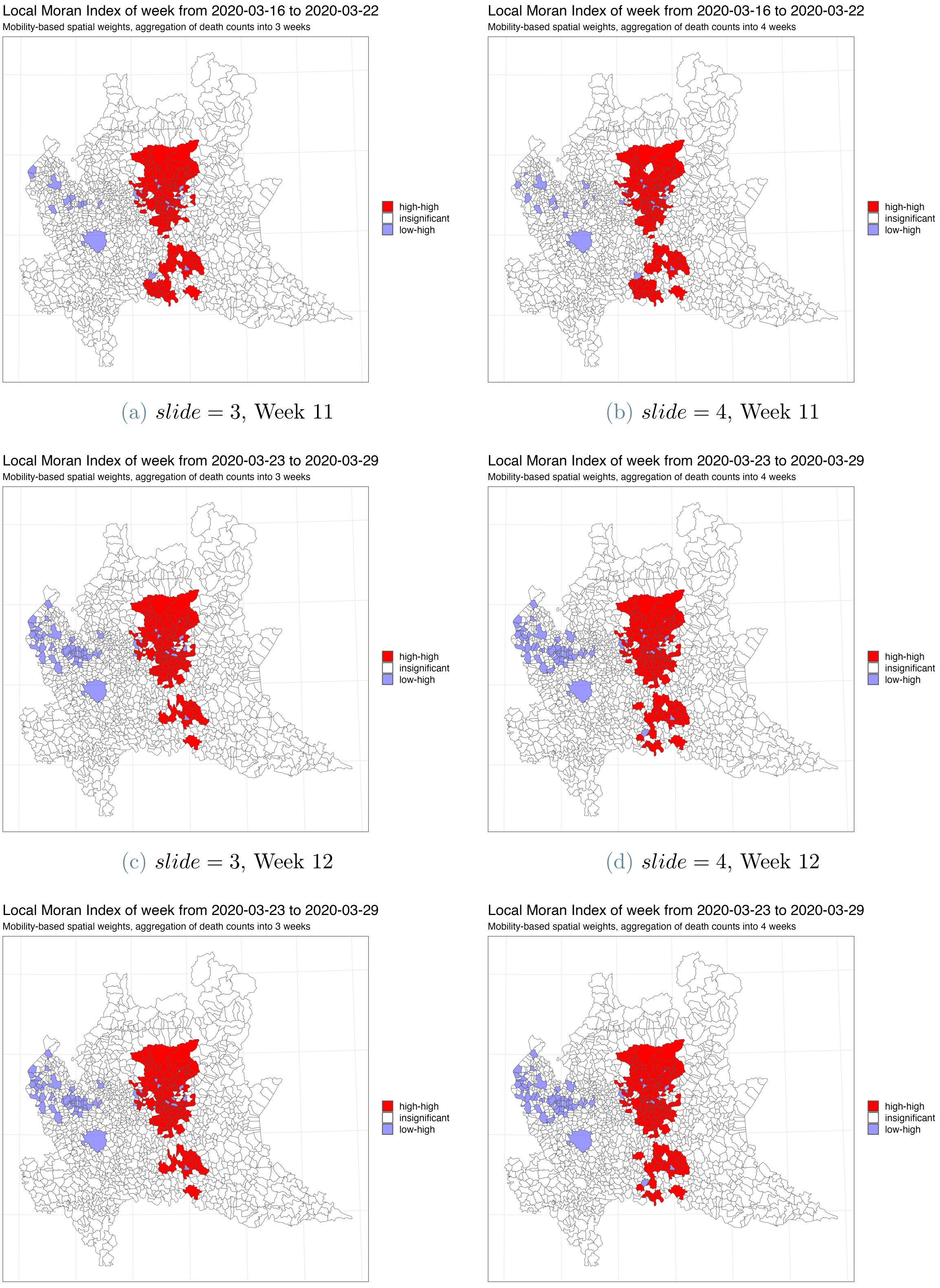
(c) $slide = 1$, Week 12

(d) $slide = 2$, Week 12

(e) $slide = 1$, Week 13

(f) $slide = 2$, Week 13

Figure A.4: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia, comparing $slide \in [1, 2]$ and $w \in [11, 12, 13]$



Local Moran Index of week from 2020-03-16 to 2020-03-22
Mobility-based spatial weights, aggregation of death counts into 3 weeks

(a) $slide = 3$, Week 11

Local Moran Index of week from 2020-03-16 to 2020-03-22
Mobility-based spatial weights, aggregation of death counts into 4 weeks

(b) $slide = 4$, Week 11

Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 3 weeks

(c) $slide = 3$, Week 12

Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 4 weeks

(d) $slide = 4$, Week 12

Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 3 weeks

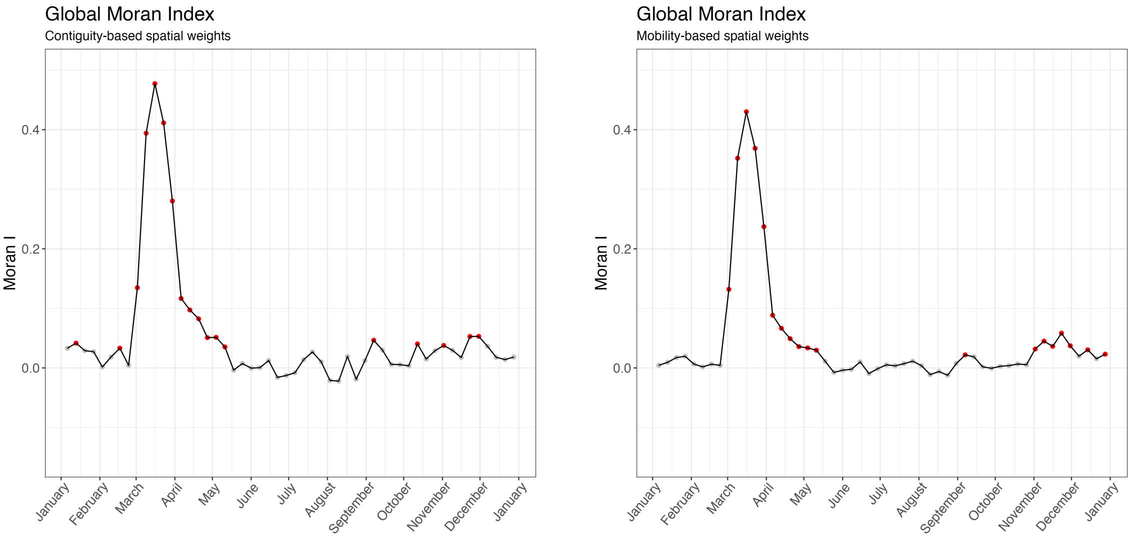
(e) $slide = 3$, Week 13

Local Moran Index of week from 2020-03-23 to 2020-03-29
Mobility-based spatial weights, aggregation of death counts into 4 weeks

(f) $slide = 4$, Week 13

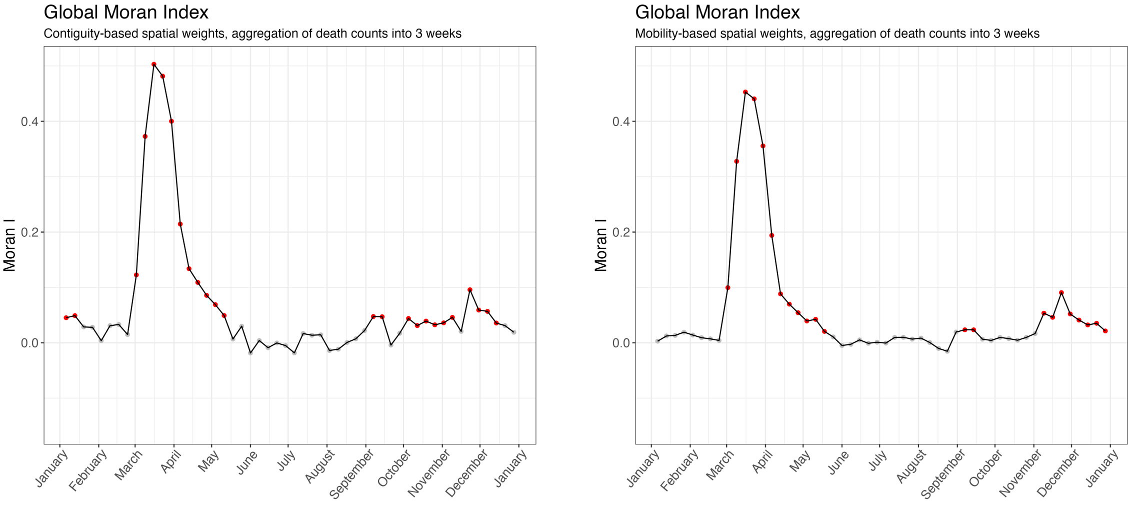
Figure A.5: Spatial clusters and outliers identified by the local Moran indexes $I_i^{[w]}$ in Lombardia, comparing $slide \in [3, 4]$ and $w \in [11, 12, 13]$

A.2. Discussion of the case $slide = 3$



(a) Contiguity-based spatial weights (b) Mobility-based spatial weights

Figure A.6: Global Moran indexes at Lombardia level based on RL mobility data and aggregation of death counts into 2 weeks



(a) Contiguity-based spatial weights (b) Mobility-based spatial weights

Figure A.7: Global Moran indexes at Lombardia level based on RL mobility data and aggregation of death counts into 3 weeks

Figure A.6 displays the curves of global Moran indexes computed with $slide = 2$ according to contiguity or mobility-based spatial weights, while Figure A.7 shows the same curves selecting *slide* $= 3$.

We can observe that there are no differences in the weeks' positive spatial autocorrelation detected during the first wave period according to the two values of *slide*. On the other hand, aggregating the death counts data into three weeks recovers a higher number of weeks showing positive spatial autocorrelation in the second wave period. Indeed, for values of $m_i^{[w]}$ computed choosing $slide = 3$, we observe 10 significant weeks for the contiguity-based spatial description (as opposed to the 5 weeks identified with $slide = 2$) and 8 significant weeks for the mobility-base case (7 for the $slide = 2$ case). The first wave period shows no major differences in the global Moran indexes, but selecting $slide = 3$ causes a significant delay in identifying interesting areas through local Moran indexes during the first wave, which is why we decide not to proceed with the aggregation into three weeks analysis. However, in case of future analysis, it could be worth exploring further the role of the time aggregation parameter.

List of Figures

List of Tables

Ringraziamenti

Vorrei ringraziare la mia relatrice, la Professoressa Francesca Ieva, e il mio co-relatore, il Professor Piercesare Secchi, per avermi dato l'opportunità di seguire questo progetto e usarlo per la mia tesi di laurea. Entrambi hanno dedicato tempo e passione al mio lavoro e mi hanno aiutata a crescere, insegnandomi come portare a termine una ricerca dall'inizio alla fine, dandomi fiducia e lasciandomi autonomia per proporre le mie idee e scelte. Grazie per avermi guidata nell'ultima parte del mio percorso universitario: ne conserverò un bel ricordo.

Vorrei ringraziare il Dottor Giovanni Chiodi e la Dottoressa Marta Galvani per avermi concesso l'opportunità di lavorare sui dati di Trenord e avermi guidata nella prima parte del lavoro con pazienza e gentilezza. Ringrazio entrambi per essersi interessati al mio lavoro e ai suoi risultati e per avermi fatto scoprire l'interesse per l'analisi dei dati di mobilità.

Un grazie sentito anche alla mia famiglia e a tutti gli amici che hanno condiviso con me ogni momento, bello o brutto, di questi cinque anni. L'Università è stata un viaggio lungo e tortuoso, ma non mi avete mai fatto mancare affetto e incoraggiamento.

Grazie a tutti,
Greta