Politecnico di Torino

ISI

ISI Foundation
& ISI Global Science
Foundation

POLITECNICO DI TORINO

Master degree in Physics of Complex Systems

Modeling the spread of COVID-19 in New York City and the impact of socio-economic inequalities

Supervisors:

Dr. Michele Tizzoni
ISI Foundation

Co-supervisor:

Prof. Alfredo Braunstein
Politecnico di Torino - DISAT

Candidate:

Claudia Deceglie

Academic year 2020/2021

Abstract

Since the start of the COVID-19 pandemic, there has been an increasing awareness about the geographical heterogeneity of the spread of SARS-CoV-2 and such heterogeneity has been observed at different spatial scales in many countries across the world. This thesis investigates the possible relations between the geographic disparities in COVID-19 incidence and the differences in socio-demographic and economic profiles of the New York City neighborhoods. Moreover, the objective of the work is also to study the effects of the interventions imposed by the NYC government on the infection spread, in the period from March to October 2020. To this aim, we use a computational approach to model the epidemic dynamics in NYC at the zip code level. A network structured meta-population model is implemented to reproduce the geographical transmission trends observed in NYC. The advantage of granular spatial modelling of the meta-population model is combined with the possibility to realistically describe the variations of mobility flows across NYC neighborhoods in the early phase of the pandemic. Mobility patterns are included in the model thanks to available mobile phone data at the census tracts level. The effects of the differential behavioral response to social-distancing policies are summarized in the model by a local risk factor, which affects the transmissibility of the infection at the spatial level in the simulations. Our results show that the simulated geographical patterns of the COVID-19 incidence correctly reproduce the actual observed spatial trend of the epidemic across NYC neighborhoods, confirming the fundamental role of human mobility and socio-economic factors in the epidemic.

Contents

Abbreviations

NYC	New York City
ZCTA	ZIP Code Tabulation Areas
UHFs	United Hospital Fund Neighborhoods
ILI	Influenza-like illness
OLS model	Ordinary least squares model
VIF	Variance Inflation Factor
SIR model	Susceptible, Infectious, or Recovered model
RD	Reaction diffusion
CBG	census block group
POI	Point of interest
ACS	American Community Survey

Chapter 1

Introduction

The novel coronavirus disease 2019 (SARS-CoV-2) emerged globally at the end of 2019 from China and has rapidly evolving into a worldwide pandemic. It has been considered the most severe public health crisis at global level since the 1918 Spanish flu pandemic, concerning its transmission and infection characteristics [1].

In recent months, many studies have been carried out in order to better understand the rapid diffusion of SARS-CoV-2, the health risks associated to it and the reasons of the differences in its spatial diffusion. Starting from Wuhan, the overall clinical severity of the infection has been studied [2], [3], [4], using publicly available data and several data-based analysis and modelling processes have been carried out to introduce estimates of infections, which are essential for the development and evaluation of public health strategies. Mathematical modeling has been crucial to understand the early transmission dynamics [5] and to evaluate the effectiveness of control measures applied in different countries. The actual observations and proposed models converge to explain how social and geographic factors are determinant in COVID-19 diffusion both at global [6] or at lower spatial scale resolution, specially in densely populated areas [6].

This work focuses mainly on New York City, given the availability of updated data from different sources and the geographical heterogeneity observed in the virus diffusion. New York City was the first epicenter of the pandemic in the United States, where a statewide stay-at-home order, the "New York State on PAUSE" executive order, was introduced on March 22, 2020 [7]. At the end of the first pandemic wave, the geographical heterogeneity associated to the infection's incidence was already documented by different studies [8].

1.1 Thesis organization

The thesis project is organized in two main parts: the first section involves the data collection, cleaning and analysis concerning the overall COVID-19 incidence after the first SARS-CoV-2 pandemic wave. Therefore, the data involved in this work refer to the period from March to October 2020. Furthermore, the socio-economic characteristics of NYC population at zip code level are investigated, in order to assess the presence of some relationships with the infection's trend.

The second main section is dedicated to the description of computational methods to study the SARS-CoV-2 diffusion dynamics across the NYC neighborhoods. The main objective of the modeling process will be to reasonably modify a basic meta-population model to introduce in the simulations aspects such as commuting patterns and risk of infection at ZCTA level.

1.1.1 Chapters main contents

The five chapters of this paper work can be summarize as follows.

The first chapter consists in the introduction of the research study.

The second chapter is dedicated to the collection of socio-demographic and economic data per zip code of NYC and the details about health data associated to COVID-19 spread. The analysis of these socio-economic data and health data and the criteria used for the data cleaning are explained and contextualized in the general research framework. The analysis is first carried out for the five principal boroughs of NYC, i.e. Manhattan (New York City County), Bronx (Bronx County), Brooklyn (Kings County), Queens (Queens County) and Staten Island (Richmond County) and then the analysis proceeds at greater geographical resolution, that is at zip codes level. Moreover, a regression analysis is proposed to study the relations between the socio-demographic fabric of NYC and the impact of the first pandemic wave.

In the third chapter, is proposed a description of the methods used at computational level to model the spreading dynamics of the SARS-CoV-2 specifically for NYC. In first place a review on the conventional metapopulation SIR model is presented to highlight the meaning and the context of the modifications that will be introduced. The proposed modifications of the model refer to the considerations given on the socio-demographic data and specific spatial heterogeneity of COVID-19 incidence in NYC. Therefore, the main objective of the computational section is to introduce factors that affect the simulations of the disease dynamics, such as commuting patters and specifics concerning the differences in the social fabric of NYC. In order to do so, commuting patters at an high resolution

spatial scale, i.e. at ZCTA level are introduced. Therefore, the collection and selection of mobility data, describing the variations in commuters' behavior during the first few months of COVID-19 pandemic will be presented.

In the fourth chapter results of simulations with the meta-population model proposed are presented and commented. The code used for the simulations is available in GitHub repository of this work [22].

In the last chapter we made some conclusions to offer to the reader a deeper explanation of the limits and the strengths of the proposed model and to suggest future improvements of this work.

Chapter 2

COVID-19 health data and socio-demographic data

2.1 Data collection at ZCTA level

2.1.1 COVID-19 NYC Health department data

The starting point of our data analysis has been to consider the overall incidence of the SARS-CoV-2 infection during the first outbreak in New York City from a geographical point of view. The spatial scale resolution that will be considered in this work is represented by the 177 ZCTAs of NYC. The ZCTAs are the ZIP Code Tabulation Areas that represents the generalized ZIP Code service areas, identifying the metropolitan area associated with mailing addresses [9].

COVID-19 incidence refers to the occurrence of new SARS-CoV-2 positive cases in NYC population over a specified period of time. Cumulative incidence data of the COVID-19 disease have been collected from NYC Health department website [10], which released detailed information since the early days of the pandemic outbreak. The data collected describe the incidence proportion [11] of the NYC population in terms of antibody test positivity to SARS-CoV-2 virus. Cumulative incidence, in fact, represents the portion of individuals residing in each ZCTA that had developed COVID-19 and had resulted positive to the antibody test in a limited period of reference. In our research work the period of time considered goes from March to October, 1st, 2020, before the formal start of the second pandemic wave in NYC. The antibody (serology) test is a specific test that can find out if a person has ever been infected by a specific virus, since it looks for the antibodies developed after the infection [12].

The maps in Figure 2.1 shows the cumulative incidence at neighborhood level, in

particular the fraction of positive confirmed cases to the antibody tests, carried out in the mentioned period, over a population unit. Test positive rate defined at ZCTA level refers to the number of people in those specific ZIP Code areas that were found positive to the antibody test per 100,000 residents. Therefore, in our study, this indicator will be identified as the main variable representing the COVID-19 impact at geographical level from the epidemiological point of view.

The first observation that flows directly from the map of Figure 2.1 is the particular spatial heterogeneity in the cumulative incidence of COVID-19 referring to the early months of the first pandemic wave. One may suspect that this heterogeneous geographical pattern is directly related to the collection of the data itself, since they may have been affected by the differences at neighborhood level in the testing ratio. In fact, as concluded in [28], testing has not been proportional to need in NYC and this is strictly dependent on the socioeconomic and racial disparities across the ZCTAs, that determine a different access to healthcare. Therefore, it might have happened that in some neighborhoods only people with mild symptoms had the possibility to take the test, especially in the early months of the pandemic. In Figure 2.2 this uneven testing ratio at spatial scale is visualized.

Despite this, the heterogeneity concerning the number of positive cases per 100,000 inhabitants at borough and also at ZIP Code spatial level is not uniquely dependent on the testing ratio, and this can be visualized in Figure 2.3.

In fact, in the scatter plot, each point on the grid represents a specific ZCTA and most of them refer to a fraction of the tested population of about $20\% - 40\%$ of the total population. Although, the same percentage of tested individuals can be associated to a percentage of people affected by COVID-19 that varies from 4% to 10%. In Queens borough, for example, given the same fraction of population tested per ZCTA, around 30%, the incidence proportion can vary from less then 4% to more than 12%.

These evidences lead us to investigate the origin of these differences in the COVID-19 spread across the neighborhoods in previews other than health sector and suggest the existence of some relationships with the social fabric of each ZCTA, characterised by specific demographic and economic indices. In the recent months, many studies with the same purpose have been carried especially for NYC: taking inspiration from one of these research works [27], it has been interesting and enlightening to analyze the demographic and socio-economic determinants of this heterogeneity in the incidence of the pandemic across neighborhoods.

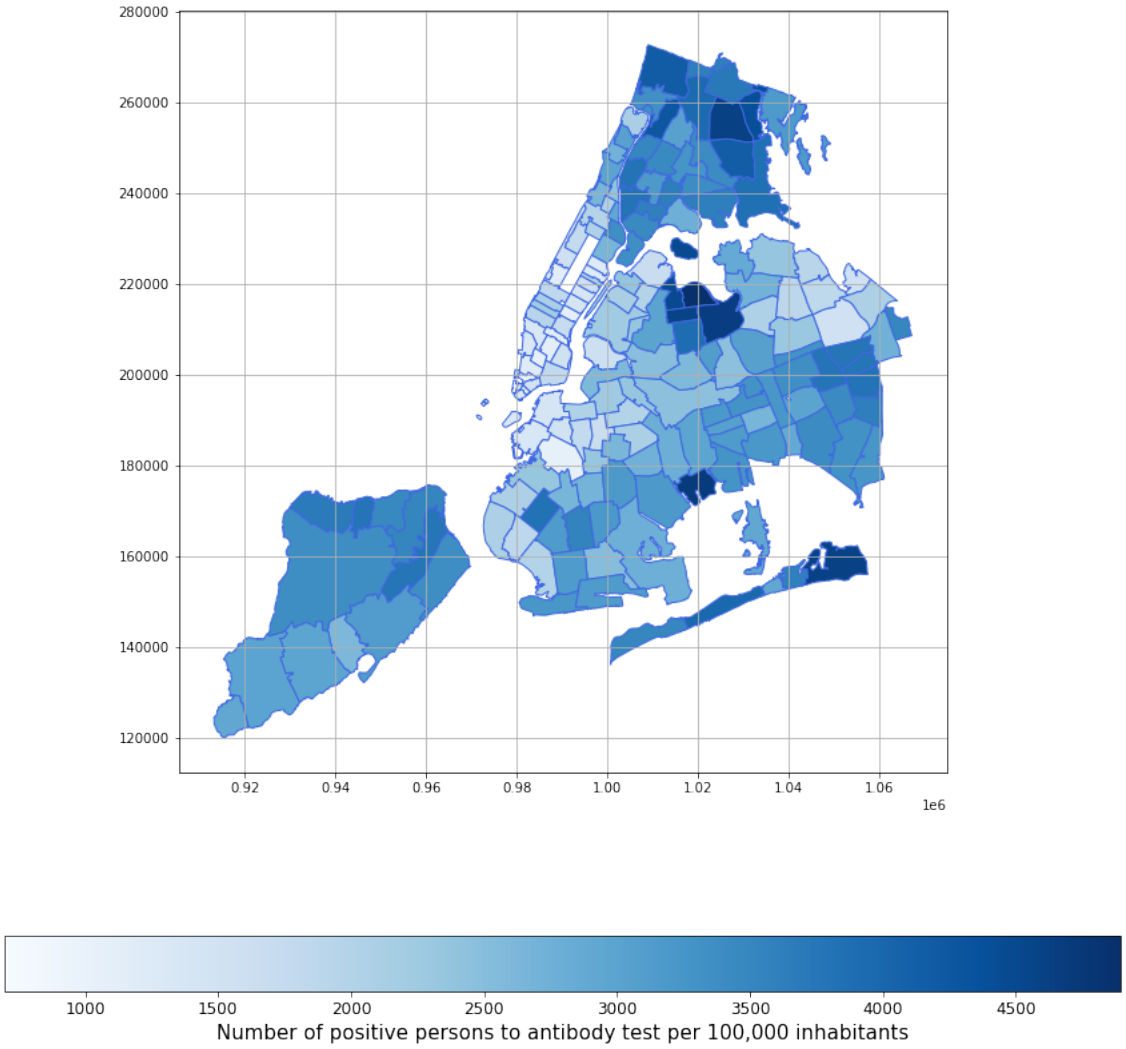

Figure 2.1: COVID-19 positive cases per 100,000 residents at ZCTA level in the limited period from March to October 2020.

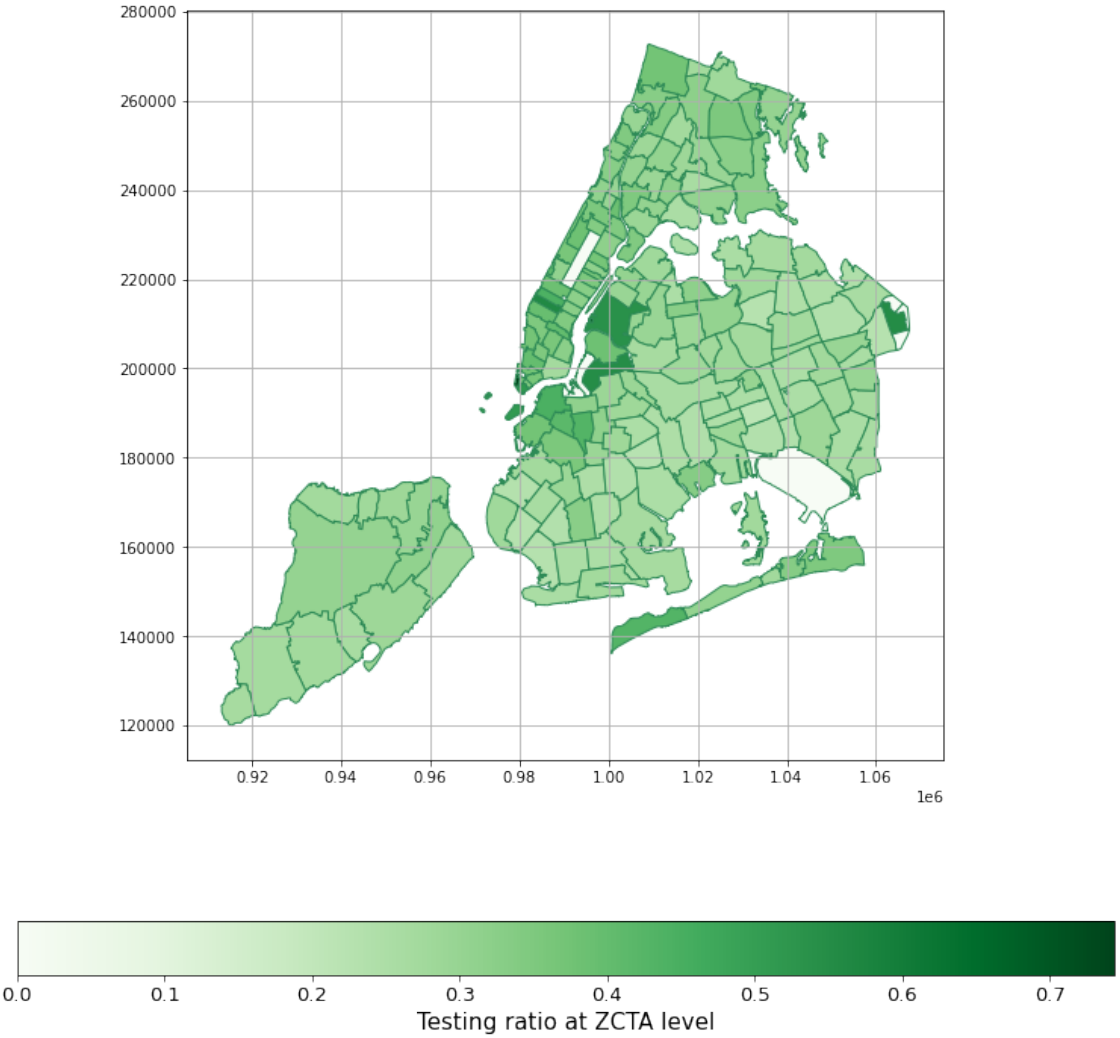

Figure 2.2: Portion of total population tested in each neighborhood.

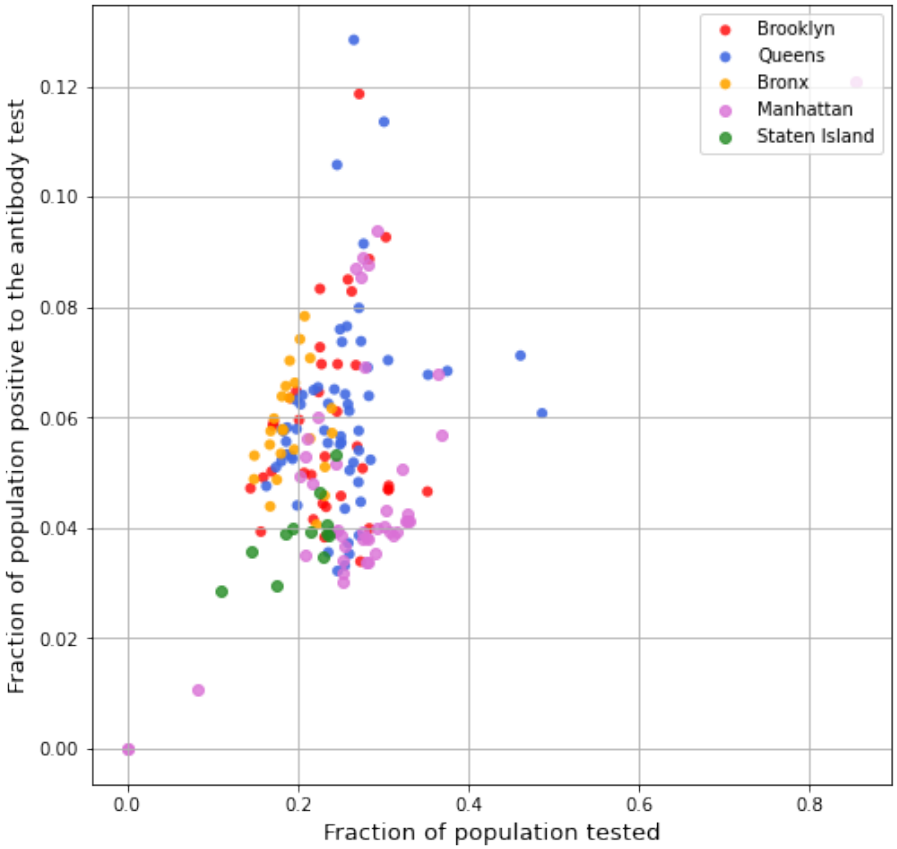

Figure 2.3: Scatter-plot of the fraction of positive tests against the proportion of tested population by neighborhood.

2.1.2 Socio-demographic data at ZCTA level

In order to perform a complete data analysis, a single data-frame unifying socio-demographic and economic data per ZCTA of New York city had been built. The data have been collected from different sources, summarized as:

- NYC Open Data website [25];
- data of New York city Health department [10];
- data sets from the 5 years American Community Survey of Census Bureau [26].

In the GitHub directory of the work [13] an R code was available to collect data from different sources. Therefore, it has been modified properly and used for our specific framework of investigation. The documentation is included in the GitHub repository of reference of this thesis project [22].

This socio-demographic data-frame has been merged with health data mentioned in the previous section, involving in particular the COVID-19 positive test rate per ZCTA, the absolute number of total positive to the antibody test and the cumulative number of tests carried out up to October 1st, 2020 per neighborhood.

2.2 Data analysis and methods

2.2.1 Correlation between socio-demographic variables and COVID-19 uneven spread across NYC neighborhoods

First, it has been important to clean and analyze the socio-demographic data, in order to select among the socio-economic determinants in the dataframe which of them can be considered as socio-economic disadvantage indicators and then investigate how they actually affect the spatial uneven virus spread across NYC ZCTAs. Therefore, an overall correlation analysis on the columns variables of the dataframe was performed, using in particular the **Pearson's Correlation Coefficient method**: it is known as the best method of measuring the statistical association between variables because it is based on the method of co-variance. In order to correctly apply this method, it's necessary to verify some assumptions:

- There should be independence of cases: observations of the different variables should be independent;
- Variables should be measured at the interval or ratio level;
- Theoretically, variables should follow a bivariate normal distribution, although in practice it is frequently accepted that simply having univariate normality in both variables is sufficient;
- There should be a linear relationship between the variables of which one wants to study the statistical association;
- There should be no significant outliers;
- Homoscedasticity of variables.

In our case, the most correct and precise indicator of the cumulative incidence of the SARS-CoV-2 infection at spatial level is represented by the percentage of individuals positive to the antibody tests per 100,000 inhabitants; so among the whole set of variables in the dataframe, visualized in the map in Figure 2.4, only socio-economic determinants that have a medium-high level of correlation with this data (*positive per 100,000*) will be considered. The Pearson's Correlation Coefficient method is based on the calculation of the correlation coefficient, that is a measure of linear correlation between two sets of data and it is calculated by taking the co-variance of the two variables and dividing it by the product of their standard deviations, as shown in equation 2.1. The correlation coefficient is expressed as a positive or negative number between -1 and 1: it gives information about the magnitude of the association, or correlation, in its value as well as the direction of the relationship between the variables involved; so values higher then |0.40/0.50| indicate a level of medium, high correlation.

$$\rho_{xy} = \frac{cov_{x,y}}{\sigma_x \sigma_y} \tag{2.1}$$

So variables with a correlation coefficients higher then |0.42| have been selected:

- *Perc_positive*, percentage of population positive to antibody tests (+0.83);
- *not_quarantined_jobs*, refers to an estimates of workers still moving to go to workplace (+0.74);
- *avg_hhold_size*, average number of individuals per household (+0.64);
- *testing_ratio*, fraction of tested residents (+0.63);

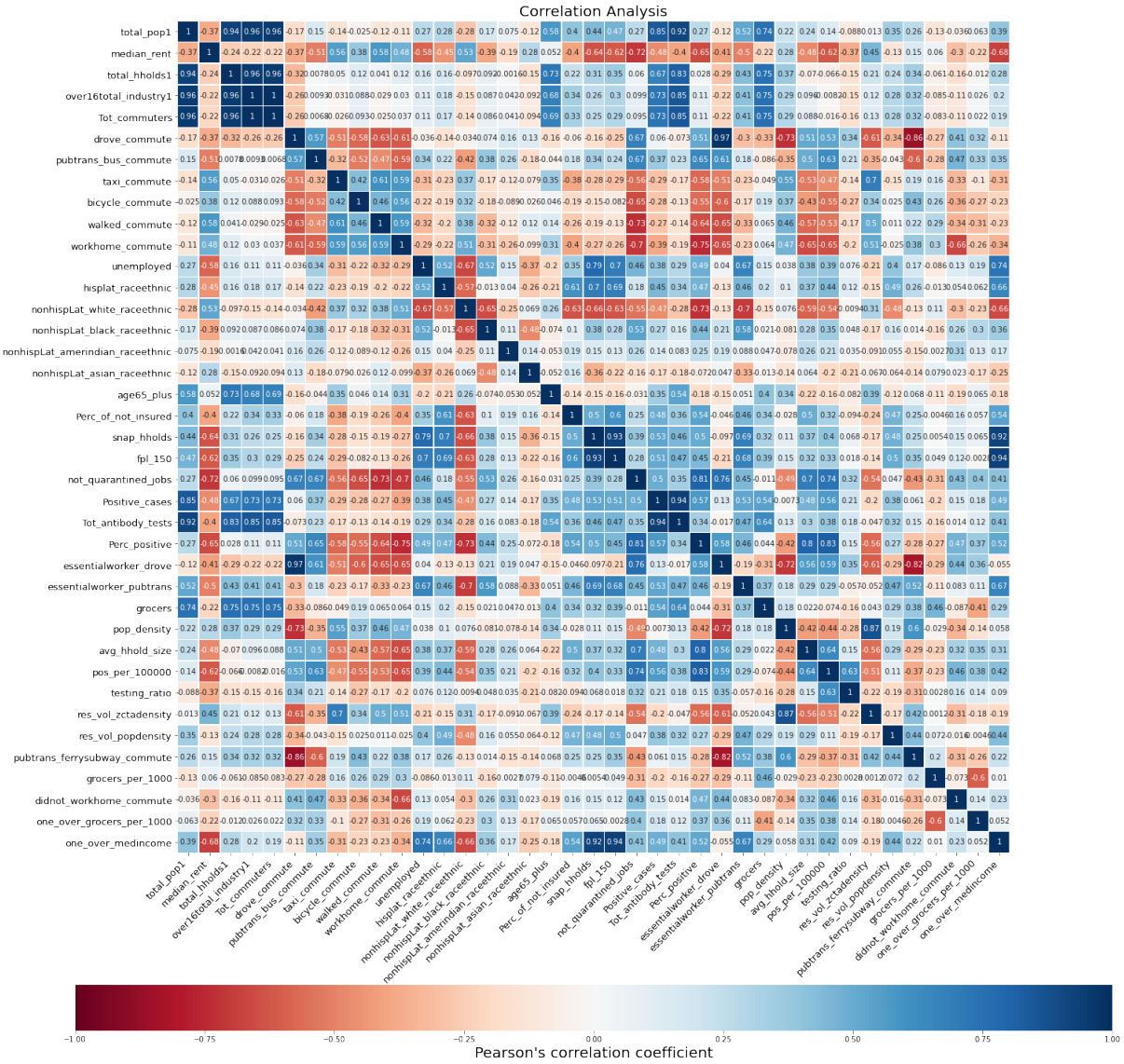

Figure 2.4: Correlation matrix for the socio-economic and health data dataframe.

- *pubtrans_bus_commute*, commuters using public transports as bus and subway (+0.63);
- *essentialworker_drove*, workers commuting by private car (+0.59);
- *Positive_cases* , absolute number of positive to antibody test (+0.56);
- *drove_commute*, commuters driving (+0.53);
- *didnot_workhome_commute*, commuters that cannot work from home (+0.46);
- *hisplat_raceethnic*, people of Hispanic and Latin ethnicity (+0.44);
- *one_over_medincome*, the reciprocal of the median income (+0.42);
- *pop_density*, population density (-0.44);
- *taxi_commute*, commuters using taxi (-0.46);
- *res_vol_zctadensity*, residents' density per zip code (-0.50);
- *walked_commute*, residents commuting on foot (-0.53);
- *nonhispLat_white_raceethnic*, people of white ethnicity non-Hispanic or Latin (-0.54);
- *bicycle_commute*, commuters by bicycle (-0.55);
- *median_rent* (-0.62);
- *workhome_commute*, commuters that work home(-0.65).

Some considerations on the variables involved in the previous list had to be done. As economic indicator the *one over median income* variable has been included [13]. This choice is reasonable given the a priori hypothesis that increased socio-economic disadvantage, such as lower median income, yields to higher infections: between the economic determinant and the indicator of the infection's incidence there is a direct positive proportionality relationship. Furthermore, the variables highly correlated with the positive case rate are represented by the mobility indicators, concerning the different type of commuting and data about workers' commuting. Relevant variables are also the ones concerning the ethnicity and this evidence is coherent with the differences in the number of COVID-19 infections and hospitalizations at the ethnic level, as evident by the data collection from the Centers for Disease Control and Prevention [29].

2.2.2 Multiple regression model for COVID-19 prevalence

Given the previous correlation analysis on the socio-demographic determinants, it would be reasonable to build up a **multiple linear regression model**, considering as dependent variable the number of positive cases per 100,000 inhabitants per ZCTA in NYC and assess which among the socio-economic and demographic variables are most relevant in determining the heterogeneous trend of the pandemic in NYC.

Multiple regression is in fact a statistical, parametric technique that uses several explanatory variables to predict the outcome of a response variable, so in this sense is the extension of **ordinary least-squares method** (OLS). The purpose of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response variable, and so to find an equation that best predicts the Y variable as a linear function of the X predictors involved, weighted by coefficients, as shown in equation (2.2).

$$Y_{exp} = a + b_1 X_1 + b_2 X_2 + ... \tag{2.2}$$

In order to perform the multiple linear regression it is necessary to verify certain assumptions with respect to the data we want to use:

1. There must exist a linear and additive relationship between independent and dependent variables. By linear dependence, it means that the two variables change at constant rate: so in this case scatter-plots can be used to show whether there is a linear relationship between our selected predictors and the outcome. By additive, it refers to the effect of X (independent variable) on Y (outcome of the prediction) is independent of the effects of the other variables.

2. There must be no correlation among variables involved as predictors. Presence of correlation in dependent variables lead to multicollinearity.

3. The error terms must possess constant variance. Absence of constant variance leads to heteroskedestacity.

4. Once done the multivariate regression, errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed: the error terms must be uncorrelated, otherwise presence of correlation in terms of error means autocorrelation and it would drastically affect the regression coefficients and standard error values.

Variables	**VIF**
nonhispLat_white_raceethnic	2.699410
one_over_medincome	2.682562
avg_hhold_size	2.489377
hisplat_raceethnic	2.088295
walked_commute	2.046059
pubtrans_bus_commute	1.921152
taxi_commute	1.911909
Positive_cases	1.637034
bicycle_commute	1.623134
didnot_workhome_commute	1.425634
testing_ratio	1.169297

Table 2.1: Socio-economic and demographic variables and the associated VIF

First of all, in order to construct the set of socio-economic and demographic variables to predict the spatial incidence of COVID-19 at zip code level, it is convenient to verify the absence of multicollinearity between determinants involved. Multicollinearity occurs when there are two or more dependent variables in a multiple regression model, which have a high correlation among themselves. When some features are highly correlated, we might have difficulty in distinguishing between their individual effects on the response variable. Multicollinearity can be detected using various techniques, one such technique being the **Variance Inflation Factor** (VIF), applied to the standardized data.

In VIF method, we pick each feature and regress it against all of the other variables. For each regression, the factor is calculated as in 2.3,

$$VIF = \frac{1}{1 - R^2} \tag{2.3}$$

where R^2 is the coefficient of determination in linear regression. Its value lies between 0 and 1: greater the value of R^2, greater is the VIF. Hence, greater VIF denotes greater correlation. This is in agreement with the fact that a higher R^2 value denotes a stronger collinearity. Generally, a VIF above 5 indicates a high multicollinearity, so the set of predictors will be composed by selected variables from the dataframe with a $VIF >= 3$, shown in table 2.1.

As it is redundant and trivially proportional to the explanatory variable, the *Posi-*

tive_cases feature is not included in the set of predictors of the parametric model.

Moreover, it has been necessary to verify, in our specific case, that the output Y, i.e. the number of positive cases per 100,000 for each ZCTA, was in a linear relationship with each socio-demographic variable involved as predictors, and this has been done using scatter-plots, as shown in Figure 2.5.

In conclusion, the predictors verifying the assumption and included in the regression model will be represented by:

- Non Hispanic/Latin white people;
- the reciprocal of the median income, social disadvantage indicator;
- the average number of individuals per household;
- number of people of Hispanic/Latin ethnicity;
- commuters on foot;
- number of commuters using public transports;
- commuters using taxi;
- commuters by bicycle;
- workers that have to commute to workplace;
- the testing ratio associated at zip code level;

The OLS Regression details are presented in table 2.2 and in table 2.3, which shows the statistical parameters associated to each predictor. The R^2 value is the statistical measure that explains how close our data are to the fitted regression line in the model. R^2 value ranges between 0% and 100% or between 0 and 1. A R^2 value of 0% reveals that the model explains no variation of the response data around its mean; however a R^2 value of 1 or 100% illustrates that the model explains all the variability of the response data around its mean. The latter R^2 value, close to 1, is mostly preferred. Generally, the higher the R^2, the better the model fits your data. Nevertheless, there are factors to consider in evaluating the validity and reliability of fit such as the sample size, the variables selection and unknown factors influencing the response.

In our case, a $R^2 = 0.848$ is sufficiently large to rely on this model and make some considerations on the P values of the variables involved, shown in table 2.3: predictors with a low P-value (< 0.05) is likely to be a meaningful addition to your model because it will indicates that the changes in this predictor's value are related to the changes in the

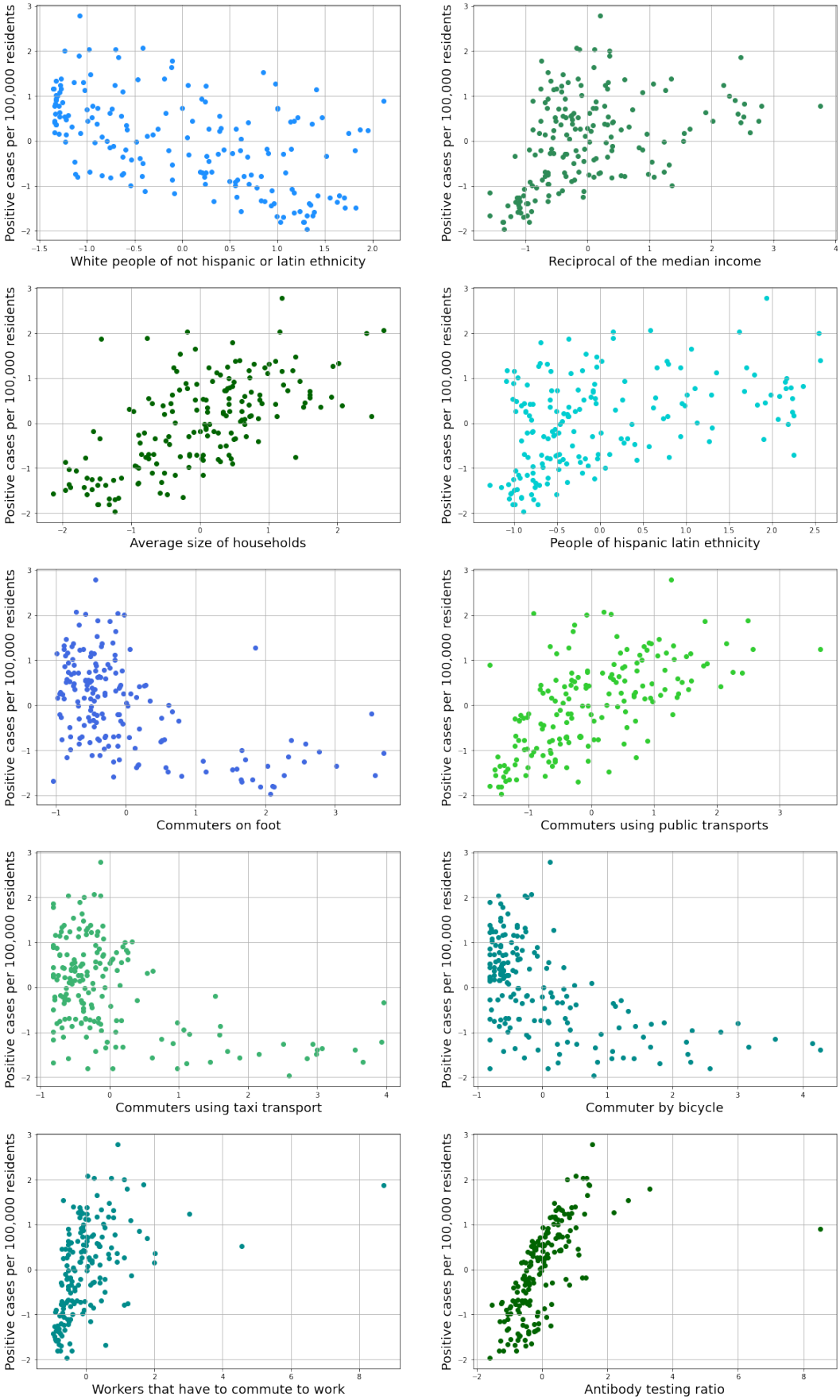

Figure 2.5: Verifying the linear relation between each predictor and explanatory variable.

Dep. Variable:	pos_per_100000
R-squared:	0.848
Model:	OLS
Adj. R-squared:	0.839
Method:	Least Squares
Log-Likelihood:	-84.155

Table 2.2: Specifications of the regression model

	coef	**std err**	**P> |t|**
const	-1.11e-16	0.030	1.000
Non-Hispanic/Latin white people	-0.1906	0.050	0.000
1/median income	-0.0325	0.048	0.502
Average size of households	0.2024	0.046	0.000
People of Hispanic/Latin ethnicity	0.1308	0.043	0.003
Commuters walking	-0.0554	0.043	0.201
Commuters using public transports	0.2081	0.042	0.000
Commuters using taxi	-0.0388	0.042	0.354
Commuters using bicycle	-0.0606	0.038	0.117
Commuters that did not work home	0.1055	0.036	0.004
Testing ratio	0.4935	0.032	0.000

Table 2.3: Multiple variate linear regression details

response variable; conversely, a larger, and so insignificant P-value suggests that changes in the predictor are not necessarily associated with changes in the response.

The variable associated to the testing ratio at ZCTA level is relevant and involved in the regression model with the highest positive coefficient than the ones associated to the other predictors and this is coherent with the reasoning done in chapter one: the counts of confirmed cases depend on how much a borough actually tests. However, this is not the only determinant in the heterogeneity audited of incidence of COVID-19 at spatial level in NYC.

In our case, almost all possible ways of commuting are involved as predictors, but the most relevant are related to the number of commuters using public transports, such as buses and subway. It has also been confirmed by studies prior to the outbreak of COVID-19 pandemic, as in paper [30], that the transmission of infectious diseases, as influenza-like illness, depends on the amount and nature of contacts between infectious and healthy individuals and this means that confined and crowded environments, such as transport hubs, can act as hot-spots for spreading disease. This is why at the declaration of the State of emergency in NYC state on March 7, 2020, government's interventions consisted in the application of voluntary precautionary measures such as avoiding public transport, following by the "stay-at-home" mandate starting the week of March 22, 2020 [31].

Therefore, in our model the average household size is a relevant parameter with a positive coefficient of linear proportionality with the output of the regression model, and this is coherent with the discussion in [13] about the effective possibility for residents of social-distancing. As moreover investigated in [32], household size and characteristics like the number of generations living together will affect the transmission of the infectious disease, household vulnerability: high population density increases the risk of rapid transmission.

The percentage of population in a specific ZCTA belonging to the Hispanic Latin ethnic group is also a representative determinant in the model. The higher is the number of Hispanic/Latin inhabitants, the larger will be the percentage of positive cases to COVID-19 infection. This result is coherent with many studies done on Hispanic/Latin communities: for example, according to the U.S. Department of Health and Human Services, Hispanic have the highest uninsured rates of any racial or ethnic group within the United States [34], and this could be determinant in control and track the disease's spread, as a larger number of residents in some neighborhoods were more likely to put off being tested for the virus, expecting it will be expensive to do so.

The significant disparities in health outcomes and COVID-19 prevalence at spatial level in NYC are then strictly correlated with the geographical pattern of economic differences and social disadvantage: in the neighborhoods where a larger number of workers has to use public transport to reach the workplace or in ZCTA where the average number of

people per household is higher, the capacity of maintaining the measures of social distance was more difficult; in the wealthiest and less densely populated districts the possibility of working from home and having more direct and simple access, from an economic point of view, to health care has determined the possibility to better track the epidemic and keep it under control in these areas.

Further research has been carried out during these months of the State of emergency to quantify this relation between the disparate impact of the epidemic and the socio-demographic differences between ZCTAs, as the social ethnicity and socioeconomic inequality, using also different regression models, as in the work by D. Carrión et al. [13].

Chapter 3

Methods and modeling

As discussed in the previous chapter, differences at geographic level in the use of public transports and, in general, the dissimilarities at ZCTA level of the commuting patterns across neighborhoods are factors that have affected the spread and the overall incidence of the SARS-CoV-2 pandemic. The evidence that the epidemic patterns in different regions are correlated with human movement and short-scale commuting patters (as workflows) has been confirmed by different studies [14], [15], [16].

Moreover, in our analysis, the differences in the socio-economic characteristics and ethnicity have been identified as determinants in the uneven spread of the virus across different zip codes.

During the year 2020, many studies have proposed a model-based analysis in order to estimate the incidence of the virus in NYC, with a particular focus on the representative factors that drove the spread of the disease, such as mobility and markers of socioeconomic status at spatial level.

In the work by Yang et al. [19], for example, the infection-fatality risk is estimated using a meta-population network model-inference system, simulating the intra and inter-neighbourhood transmission of SARS-CoV-2.

Similarly, in our study we propose a computational approach to reproduce the NYC spatial heterogeneous pattern of cumulative incidence of the SARS-CoV-2 disease.

The methodology consists in building a metapopulation model which simulates the dynamics of the virus diffusion on a network. This is done by integrating in a basic model the effect of the mobility variations during the early months of the COVID-19 pandemic in order to assess how the differences at zip code level affect the simulations of such dynamics.

3.1 Modeling the spatial spread of infectious diseases

The mathematical and computational modeling of infectious diseases is key to study the mechanisms by which infectious diseases spread, to predict the future course of an outbreak and can be determinant in the evaluation of strategies to control an epidemic.

In particular, in the case of the COVID-19 pandemic, one of the distinctive aspects of the spread of the disease is its spatial diffusion and the concurrent role of human daily commuting and mobility patterns, as proved by the previous data analysis.

As pointed out in previous studies, such as [35], network structures have emerged as a powerful framework to incorporate mobility patterns within the mathematical modeling of epidemics. Networks have been extensively used in predicting the spread of infectious diseases where individuals interact with a limited set of others, defining the graph through which the ILI can spread [17], [18].

Meta-population models are largely used in computational epidemiology [20], [21], [19], [35]. They are a type of spatial model which investigate interactions and movements among different sub-populations of the same species, across time and space. It is an extension of more conventional population-level compartment models that typically assume homogeneous mixing and implicit interactions within a population.

The objective is to simulate the dynamics of the transmission of an ILI infection in each sub-population, i.e. in each node of the network under study, whose dimensions depend on the spatial resolution required. In our case, the considered network is formed by the 177 nodes, i.e. the 177 ZCTAs of NYC. The commuting patterns between neighborhoods will be introduced as flows on the edges of the graph. The mobility variations within each ZCTA will not be taken into account.

3.2 Epidemic meta-population model

The basic epidemic meta-population model is represented by a reaction–diffusion (RD) meta-population model [36], where the whole population is divided in sub-populations and each of them is classified with respect to their role in the epidemiological process. Each sub-population is in connection with the others by the commuting flows of individuals moving across the spatial structured network, as shown in figure 3.1.

The transmission mechanism of the virus takes place in each node, simulating the infection spread between individuals in the same ZCTA and it is described by a **SIR compartmental model**, in which each individual can be either susceptible, healthy (S), infectious (I) or recovered/removed (R).

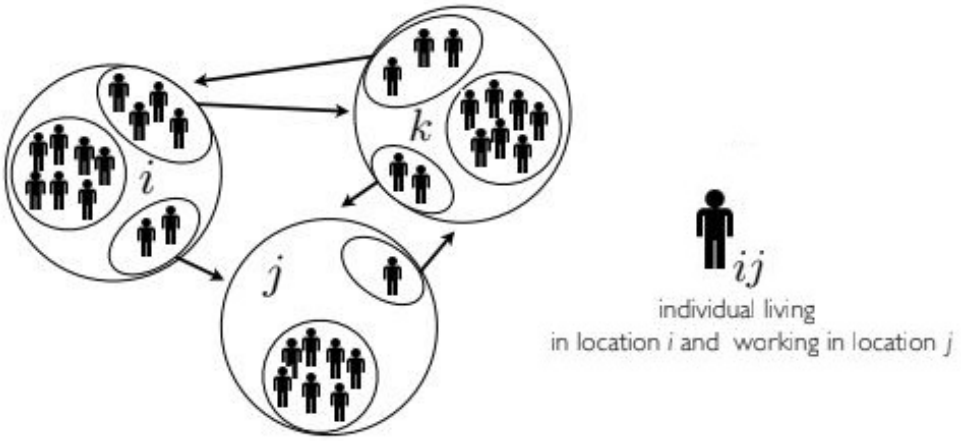

Figure 3.1: Spatial scheme of the meta-population model with commuting flows

Figure 3.2: SIR model: state transitions

Form a theoretical point of view, the classical SIR model is a compartmental model without vital dynamics, i.e. without considering the birth and death of individuals, since the timescale for modeling is much smaller than the average population turnover. The recovered individuals on the network are assumed immune to the disease once they have recovered and no more contagious. In scheme 3.2 all possible individuals' state transitions in the SIR model are summarized.

In its continuous version, this model can be described by the ODE system 3.1:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta IS}{N}; \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \mu I; \\ \frac{dR}{dt} &= \mu I \end{aligned} \tag{3.1}$$

and the normalization condition 3.2, where N represents the total population on the

network, must be verified.

$$S(t) + I(t) + R(t) = N \tag{3.2}$$

The fundamental parameters involved in model are:

- β, the transmissibility of the infection, corresponding to the reciprocal of the typical mutual contact time between individuals;
- μ, rate of healing, or the reciprocal of the average number of days after which the subject is no longer contagious and is removed.

The dynamics of the infectious class depends on the ratio 3.3, defined as the basic reproduction number:

$$R_0 = \frac{\beta}{\mu} \tag{3.3}$$

The RD process involved in the simulation model is the time-discrete version of the SIR model. The dynamics of the process is separated into two distinguished moments: the work time, when commuters move towards their working districts, and the home time, when they are assumed to be back home. Therefore, the compartmental matrices, representing the possible states of the individuals on the network, are updated during the simulation. The single elements of these matrices will indicate the number of susceptible, infected and recovered who live in neighborhood i and work in district j for every pair of i, j as S_{ij}, I_{ij} and R_{ij}, respectively.

The individuals on the network can be infected in their workplace during work time and then spread the disease once they travel back home during home time or vice versa. Infectious commuters are allowed to move on the network. In the basic model, for the sake of simplicity, the behavioral changes associated to the possible severity of clinical symptoms are not considered. Each day of a simulation is considered as a typical working day, no weekends or holidays are introduced.

The basic simulation algorithm works as described in support information section in the study of Tizzoni et al. [38].

At the starting point of the simulation, each node of the network is initialized with its resident population $N_i = \sum_i Nij$, where N_{ij} is the matrix element that refers to the number of individuals who live in ZCTA i and work in ZCTA j.
Individuals are labeled according to their health status and divided into the three compartments: S_{ij}, I_{ij} and R_{ij}. The entries of the compartmental matrices S, I and R are initialized by setting $S_{ij} = N_{ij}$ for every combination ij except for the initial seeds. Therefore, at the beginning of the simulation the only non-zero entry of the matrix I

will be $I_{ss} = 10$. This means that in a specific neighborhood, defined *seed*, there are 10 individuals already affected by the disease.

Each simulation time step represents a workday, divided in two moments, work time and home time, assumed to have equal length, 12 hours each. During work time, the force of infection in node i is calculated as 3.4:

$$f_i{}^{work} = \frac{\beta}{2} \frac{I_{ii} + \sum_j I_{ji}}{N_{ii} + \sum_j N_{ji}} \tag{3.4}$$

where β is the daily transmissibility and the factor $1/2$ takes into account that we are considering half a day. The number of new infected individuals among those who work in i is extracted using a random binomial sampling 3.5:

$$\Delta(S_{ij} -> I_{ij}) = Binomial(S_{ij}, f_i{}^{work}) \tag{3.5}$$

Analogously, during home time, the force of infection in node $i$is calculated as 3.6:

$$f_i{}^{home} = \frac{\beta}{2} \frac{I_{ii} + \sum_j I_{ij}}{N_{ii} + \sum_j N_{ij}} \tag{3.6}$$

New infected individuals among those who live in i ZCTA are extracted using a random binomial sampling, but now using the force of infection at home 3.7:

$$\Delta(S_{ij} -> I_{ij}) = Binomial(S_{ij}, f_i{}^{home}) \tag{3.7}$$

Recovery transitions happen both during home and work time, as spontaneous process, with constant probability $\mu/2$, depending on the defined recovery rate.

3.3 Meta-population model including effects of non-pharmaceutical interventions

Given the basic meta-population model, the purpose of this work is to propose some modifications of the simulation algorithm, in order to include the real impact of non-pharmaceutical interventions that had been implemented in NYC during the year 2020.

3.3.1 Commuting patterns

First, we would like to introduce the behavioral changes in the mobility patterns due to the "stay-at-home" order and the declaration of the state of emergency on March 22, 2020. The commuting in NYC has been affected also in different ways at geographical level, due to the socio-demographic determinants discussed in previous chapters, such as the effective access to testing, the possibility of use private means of transport instead of public ones, the ability to work remotely and the number of persons per household. At the outbreak of the pandemic, for example, as described in article [33], people living in Manhattan borough and in the NYC richest neighborhoods had the possibility to leave their houses to vacation places or to work from home. These determinants have affected the trends of commuting patterns in NYC in the following months, with the effect of the increased possibility of social distancing only for some communities.

In general, an high spatial and temporal resolution of mobility flows at different geographic scales over time may help to monitor the epidemic spreading dynamics and deepen our understanding of the actual human responses under the public health crisis.

In order to introduce these data in the modeling, it has been crucial to better understand from a quantitative point of view the variations in commuting flows in NYC during the first months of SARS-CoV-2 pandemic, i.e. the period of reference of this work, March - May 2020.

Mobility data at borough level and then at ZCTA level have been collected from the GitHub repository available from the research work [39].

Mobility data description

The above mentioned data set have been built up by analyzing anonymous mobile phone users' visit trajectories to various places provided by SafeGraph. Mobile phone data have been used in different studies [40], [41], [42] to predict the spatial spread of an epidemics

and represent faithfully the commuting links between highly connected locations [36]. In order to give data a structure, the daily and weekly dynamics origin-to-destination (O-D) population flows are computed at the different geographic scales: at county and census tracts level.

The place visitor patterns are retrieved from the SafeGraph COVID-19 Data Consortium, by gathering millions of anonymous GPS pings for United States, collected from numerous mobile applications tracked and cleaned to remove noise. Users' home places are estimated and aggregated at the level of census block group (CBG). Then, those users' visits from home places to points of interest (POIs) are tracked. POIs are the primary venue for tracking place foot-traffic by SafeGraph. The home place of a user refers to the place where he/she has spent most of his/her night time during the last six weeks. For each day, GPS pings of each device are clustered and only those clusters during night time hours (6pm - 7am local time) are kept. Therefore, the most frequent CBG over the last six weeks that reflects the primary night time location is used as the "home location" for each user. By aggregating home places to CBGs, user privacy can be protected, and no individual records can be traced.

Active users' visits to POIs are produced with a similar strategy. By using several clustering methods, such as density-based spatial clustering for applications with noise, GPS pings are grouped together, so that each cluster contains a set of potential POIs and associates with CBGs. The best place for a given cluster is classified by performing machine learning methods involving several entangled features. Thereby, each user's visits from home place to various POIs and CBGs are identified.

From these data, visitor flows are calculated at different scales. The two major human mobility flow metrics are denoted as daily CBG to CBG visitor flows and weekly CBG to POI visitor flows. In the daily CBG to CBG visitor flows metric, each row contains an origin CBG and a destination CBG, as well as the number of mobile phone-based visitor flows from the origin CBG to the destination CBG. Every day, the number of unique mobile phone users who live in the origin CBG and visits to the destination CBG are recorded, clustering GPS pings and involving only those clusters (i.e., not a single trajectory) with a duration of at least one minute. In this way, the daily mobile phone-based visitor flows between CBG and CBG are grouped and summed up. For the weekly CBG to POI visitor flows metric, a mapping of CBGs to POIs is provided. In other words, the number of unique visitors who live inside the origin CBG and visit the destination POI in one week are counted.

The two mobile phone-based visitor flows metrics are both processed at the CBG scale. Then, all data are further aggregated into three different spatial scales: census tract, county, and state. Providing a multi-scale flow data set allows to have a more comprehensive view of human mobility and spatial interaction patterns. Furthermore, the O-D (origin - destination) flow data set is generated at the three geographical scales,

respectively, by assign to census tract, county, and state's geographically unique identifier to each origin CBG and destination CBG to which they belong.

The calculation of visitor flows at the three spatial scales is based on mobile phone users detected by SafeGraph, not on the entire population. As discussed in work [39], these users account approximately for 10% of the entire population in the U.S. However, studies as [43] have shown that a good representative sample of the entire population can reflect general human mobility patterns. Therefore, to infer the short-term dynamic mobility flows during the COVID-19 pandemic, the official American Community Survey [26] population data with mobile phone visitor patterns has been used. The inference has been realized using the following equation 3.8:

$$pop\_flows(O, D) = visitor\_flows \ * \ \frac{population(O)}{num\_devices(O)} \tag{3.8}$$

where *pop_flows* is the estimated dynamic population flows from geographic unit O to geographic unit D, *visitor_flows* is the computed mobile phone-based visitor flow from O to D, *population (O)* indicates the population at the geographic unit O extracted from the ACS, and *num_devices(O)* refers to the number of mobile devices residing in O.

Analysis of mobility data

Therefore, given this framework, the data set considered in our work provides the weekly variations of mobility flows at county to county level. In the figure 3.3 a temporal analysis on commuting flows between NYC's counties from February 24, 2020 to May 3, 2020 is shown: in particular, data presented are normalized with respect to the first week's number of commuters of each borough, in order to better visualize the differences in decreasing or increasing of the number of trips to and from the NYC's boroughs and the variation of the internal commuting in each district.

From the graphics, it is evident that journeys towards Manhattan district have significantly decreased compared to those to other boroughs: this could probably be linked to the possibility of citizens to respect the government measures imposed and to operate smart-working according to the employment sector. In addition, as explained above, many Manhattan's residents could effectively leave their houses given the availability of vacation places outside the district. The neighborhoods where internal commuting has decreased the fewer are Bronx and Brooklyn: around the 30/40%, while in Manhattan the average number of daily internal journeys has dropped by up to 55%.

Since our geographical resolution of interest is at ZCTAs level, some further managing of the data collected from the GitHub repository of [39] study has to be done. As described before, the mobility flows' data were presented at three different geographical scale. Therefore, in order to use the data at ZCTAs level, population flows at census tracts level have been considered and summed up properly. The methodology used is properly described in the R script *"geoidtozcta_nyc_script"*, uploaded in the GitHub repository of this work [22].

Moreover, since data were given as weekly flows, it has been necessary to take into account the temporal scale of simulation. Each step of the simulation algorithm corresponds to a single work day, then the number of weekly population flows at ZCTA level has been averaged on a week, in order to have approximately the number of daily journeys of type origin-destination.

In the basic version of the model, the number of commuters on the network, from each node to the others, was fixed and given as input.

In our case, the variations of the commuting patterns are introduced at each week of the simulation, instead of modifying the absolute number of commuters. The variations introduced represent the relative decrease or increase of the number of the O-D journeys between ZCTAs compared to the value of the first week of the simulation. The equation 3.9 shows how these relative variations, i.e. the weight w_ij associated to each edge of the graph, are calculated for each week. These data concerning the relative variations of commuting patterns will be given as input files of the code. Therefore, the number of commuters effectively moving on the network is updated at each week of the simulation. The number of susceptible commuters varies according to the relative weight w_{ij}, as shown in equation 3.10, and the same operation is performed for the other compartmental matrices (infected and recovered individuals).

$$w_{ij}^{t+1} = \frac{pop\_flows_{ij}^{t+1}}{pop\_flows_{ij}^{t}} \tag{3.9}$$

$$S_{ij}^{t+1} = S_{ij}^{t} * w_{ij}^{t+1} \quad \forall i \neq j \tag{3.10}$$

3.3.2 Parameters evaluation: Risk Index at ZCTA level

The R_t parameter has been updated at each week of the simulation to better reproduce the progress of the pandemic in NYC. This has been done according to the overall trend recorded for NYC during the early months of the pandemic. When the whole population is susceptible and no interventions are in place, the R_t coincide to the basic reproductive number R_0 and it reflects the transmissibility of an infection in that population. In case

Period	R_0/R_t
24/02/2020 - 01/03/2020	3
02/03/2020 - 08/03/2020	2.99
09/03/2020 - 22/03/2020	2.22
23/03/2020 - 29/03/2020	1.37
30/03/2020 - 12/04/2020	0.93
13/04/2020 till the end of simulation	0.56

Table 3.1: Estimates of R_0 and R_t parameters introduced in the simulation algorithm

some interventions are introduced, it could be fundamental to evaluate the effectiveness of the measures on the transmissibility of the disease. However, it may be very difficult to distinguish the impact of a single intervention. In the work [44] some estimates on the R_t value of NYC have been done and we introduced them in the model to simulate the framework of the first ten weeks of the pandemic. We could summarize the associated value of the reproductive number R_0 and then the R_t values in the simulations, as shown in table 3.1:

In the first days following the declaration of the first confirmed case of COVID-19 in NYC, the R_0 was very high. It decreased during the next two weeks, when NY State declared a state of emergency and public awareness and voluntary precautionary measures (e.g. avoiding public transit) increased. Following the stay-at-home mandate starting the week of March 22, R_t dropped substantially, until dropping below zero during April, 2020, when face covering in public places was implemented as control measure. Therefore, in out model we try to reproduce this trend manually updating the value of the reproduction number.

Moreover, in the model, an indicator of the socio-economic disadvantage at geographical level has been introduced, given the analysis done in Chapter 2. The inability to socially isolate could refer to the usage of the public transports and contribute to greater the exposure risks among communities. In the study [13] MTA transit data have been used to define a risk index at United Hospital Fund (UHF) neighborhoods level for NYC. UHF neighborhoods are composed of adjacent ZCTAs approximating community districts.

Therefore, in order to introduce this risk factor in our model, it has been defined at ZCTA level according to the geographical correspondences.

This factor will be multiplied by the β parameter during the simulation and will be equal to 1 for neighborhoods at high risk of infection and 0.5 in case of ZCTAs at low risk of infection with respect to the socio-demographic framework. Moreover, the μ parameter in the code is fixed and equal to 1/3, since we are assuming that the average number of days to recover, after which the individuals are no more contagious is approximately three days. This assumption will not affect significantly the results of simulations given by such a simple model. The map of NYC's ZCTAs and the associated risk index is shown in Figure 3.4.

3.3.3 Pseudo algorithm

Here is presented the pseudo-algorithm of the model, introducing the commuting variations and the update of the R_t and β factor. In each simulation, the city seed, i.e. the node in which first infected subjects are placed, is randomly selected.

Algorithm 1: Simulation algorithm

read database of total population per ZCTA
setting initial commuting patterns
initialize the risk index associated to each node of the network
set R_0 with its initial value

for *all run r* **do**
  for *all timesteps t* **do**
    if *week changes* **then**
      update the R_0 value
      calculate the new β factor
    end
    Work time:
    for *all nodes i* **do**
      evaluate the force of infection at work
      extract infectious using random binomials with probability f_i^{work}
      extract recoveries using random binomials with probability $\mu/2$
      update compartmental matrices
    end
    Home time:
    for *all nodes i* **do**
      evaluate the force of infection at home extract transitions using random binomials with probability f_i^{home}
      extract recoveries using random binomials with probability $\mu/2$
      update compartmental matrices
    end
  end
end

Figure 3.3: Variations in commuting patterns from March to May 2020

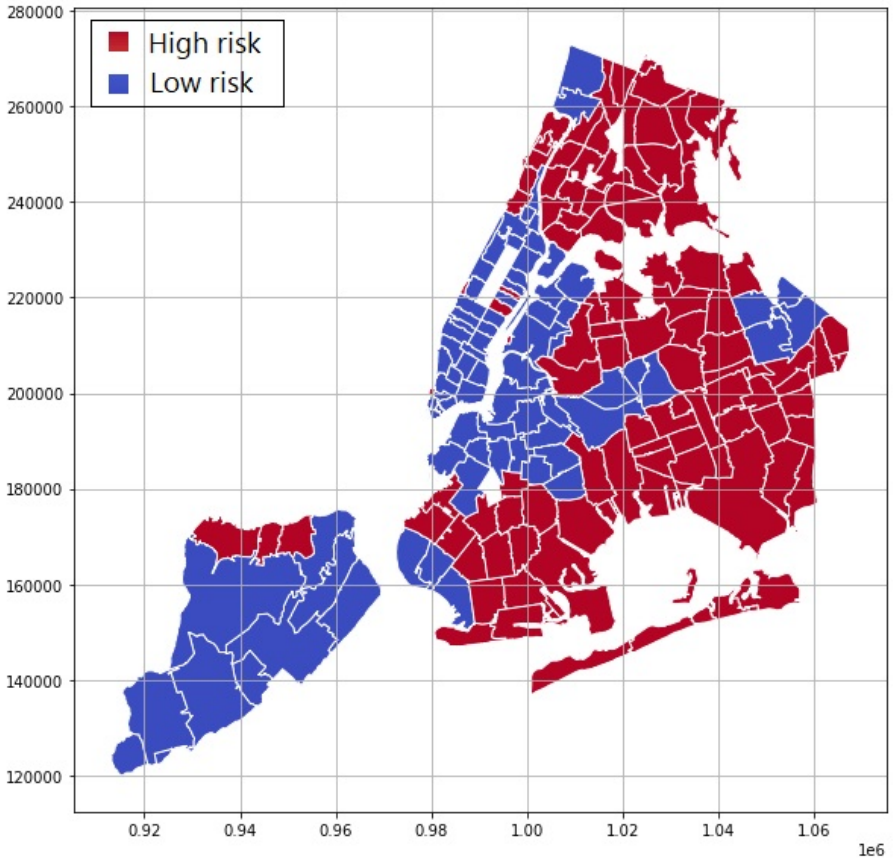

Figure 3.4: NYC's risk index at ZCTA level.

Chapter 4

Results of simulations

4.1 Geographical patterns and numerical analysis

The meta-population model described in the previous chapter has been used to simulate the dynamics of the COVID-19 spreading in the early months of the pandemic's outbreak. The results of each simulation of an epidemic spread are collected and manipulate, using a R script, given the GitHub repository of this work [22]. In Figure 4.1 we can see the median proportion of recovered individuals over total population for each ZCTA resulting from a set of one hundred simulations. Whereas, in Figure 4.2 the actual cumulative incidence of COVID-19 is displayed at ZCTA level.

The similarity between the spatial infection patterns of the two maps is evident: geographical differences in infection incidence in simulations reproduce the observed dishomogeneous trend of infection in NYC, across the different neighborhoods. However, from a numerical point of view, the results inferred from the simulations differ from the actual data for all ZCTAs by an order of magnitude. This conclusion will be explained further in the following sections, where a detailed analysis of all the factors involved in this modeling process is presented.

4.2 Statistical analysis of the results

We present the statistical analysis of the simulations results. The distribution of the numerical data concerning the fraction of population recovered after one hundred simulations with our meta-population model is shown in Figure 4.3. Data have been aggregated from ZCTA level to borough to visualize the statistical distributions of simulated data across

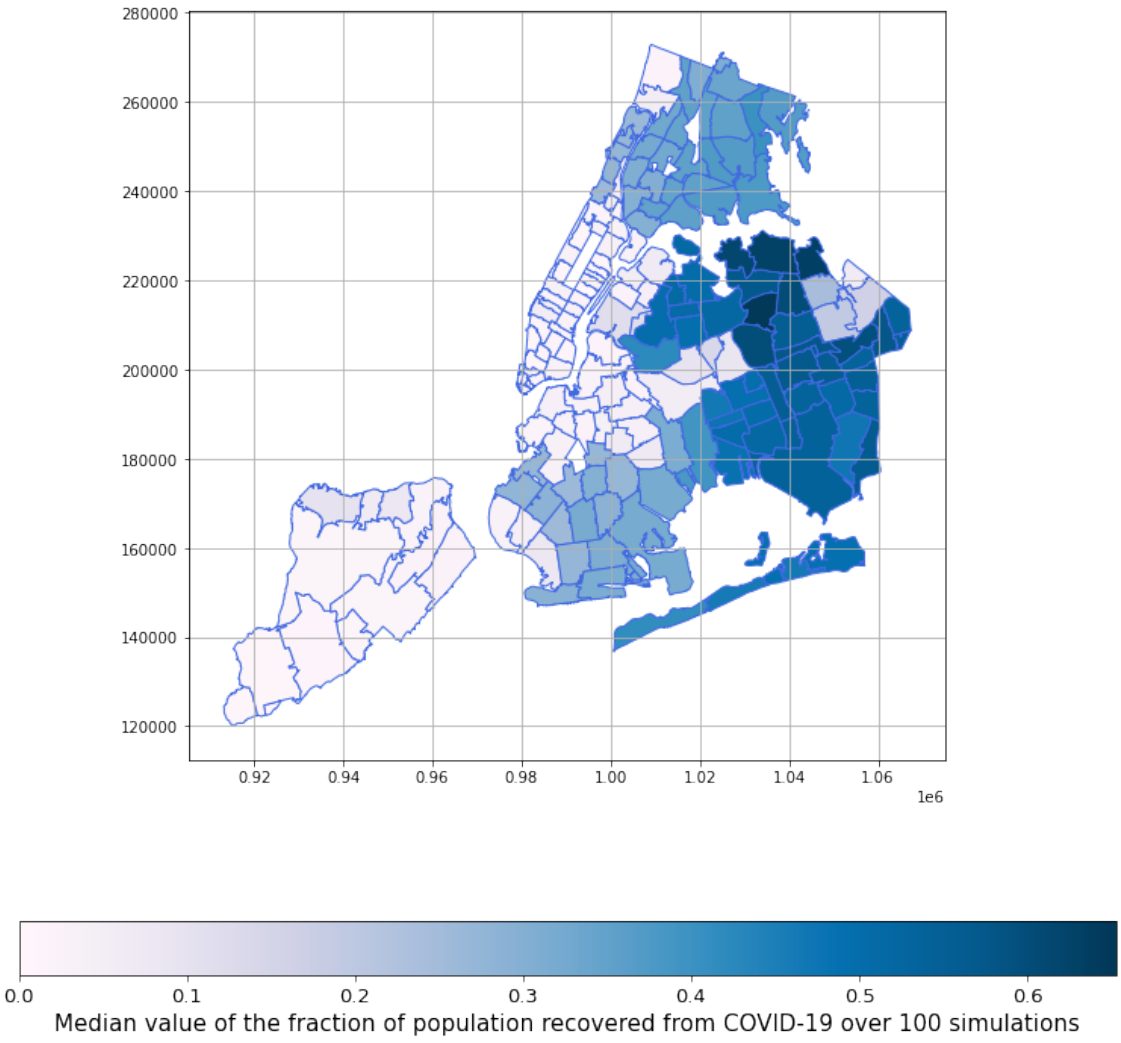

Figure 4.1: Results of simulations using meta-population model, including commuting patterns and risk index.

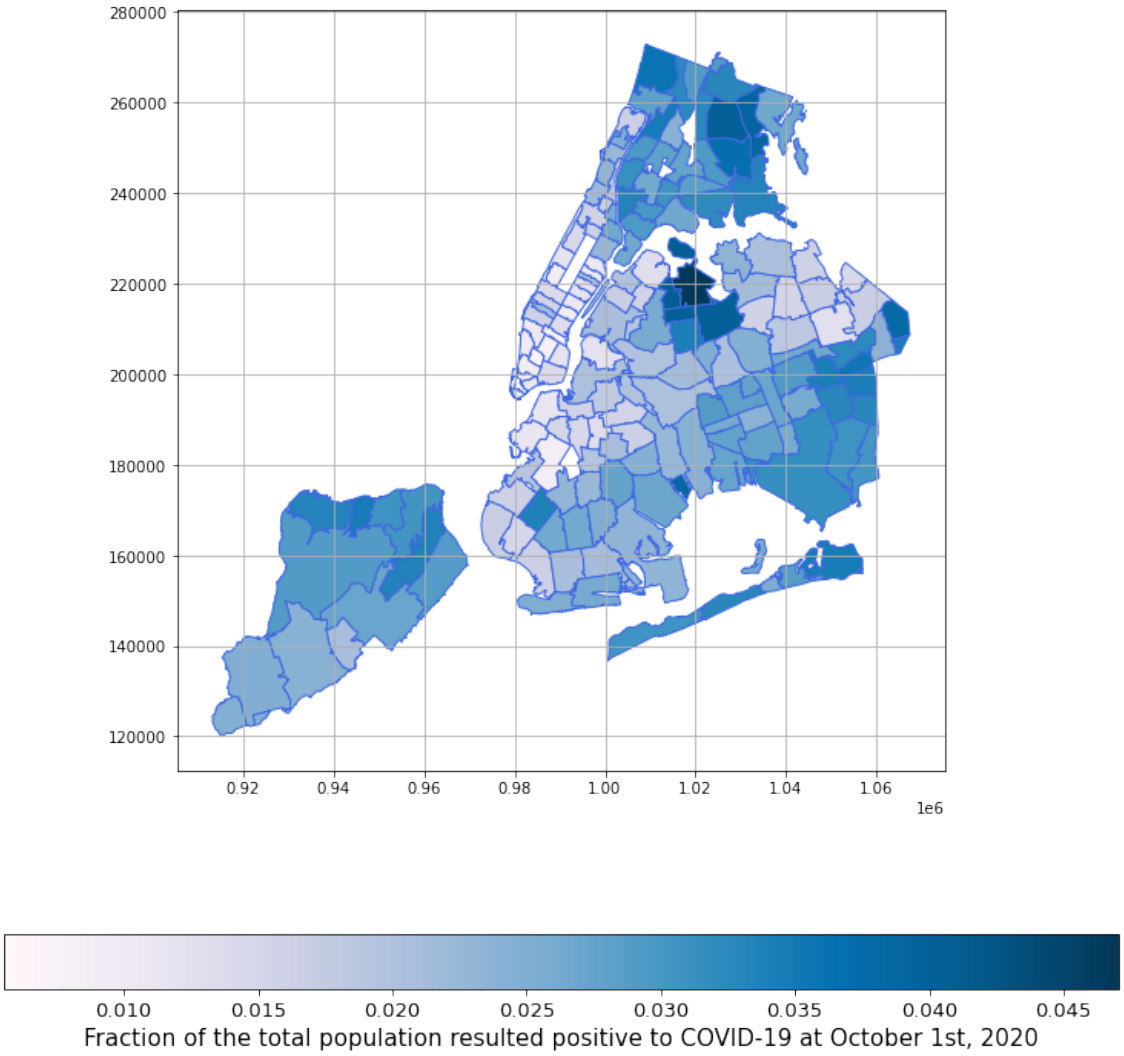

Figure 4.2: Actual cumulative incidence of COVID-19.

the five borough of NYC. As displayed in figure, the distributions of the median value representing the portion of recovered population associated to Manhattan and Staten Island are compact around the median value. Conversely, the distributions associated to Bronx, Brooklyn and Queens borough are asymmetric and long-tailed distributions. This evidence can indicate a greater heterogeneity in the attack rate of simulations across the ZCTAs of these NYC counties.

In Figure 4.4, we plot simulated over actual data indicating the portion of individuals infected over the total population in logarithmic scale. The logarithmic transformation is performed to overcome the visualization limits due to the differences in order of magnitude of the numerical results. From the Figure 4.4 we notice that except for some Manhattan and Staten Island ZCTAs the predicted fractions of recovered individuals are overestimated. In particular, Bronx, Queens and Brooklyn are the boroughs for which we visualize the largest prediction errors.

4.3 Discussion

Although the simulations correctly reproduce the geographical trends of the infection spread, the median number of recovered individuals is in general higher compared to the actual data. The simulated fraction of recovered individuals over the total population ranges between 0% and 60%, while the actual data for the same variable ranges between 0$ and 5%. This discrepancy can be explained by several factors, related to the data used in the modeling process and the limits of the model itself.
First, mobility data are inferred from a sample of individuals. Safe Graph provides mobile data tracking millions of anonymous mobile phone users' in the US and the data used in the model are the result of a data manipulation process, explained in details in chapter 3. As explained in the study [39], visitors duplication is certainly present in the mobility data set. This may have been determined by the aggregation of the commuting flows at different spatial scale resolutions, determining the overestimation of population flows between ZCTAs. Data bias is a common issue for large-scale mobile phone data and may influence the representativeness of the data set. In fact, the dynamic mobility flows are inferred from mobile phone applications by users, knowing that smartphone applications usage is not the same across the different age groups. In addition, an other limit that can be recall concern the identification of home and work locations as the only two POIs in the manipulation of the mobile phone data [38]. Furthermore, one single source of mobility data cannot provide a comprehensive description of human movements across all spatio-temporal scales [38] that can be relevant for a specific disease transmission. Although, focusing uniquely on one-type human movement, that is daily commuting, has

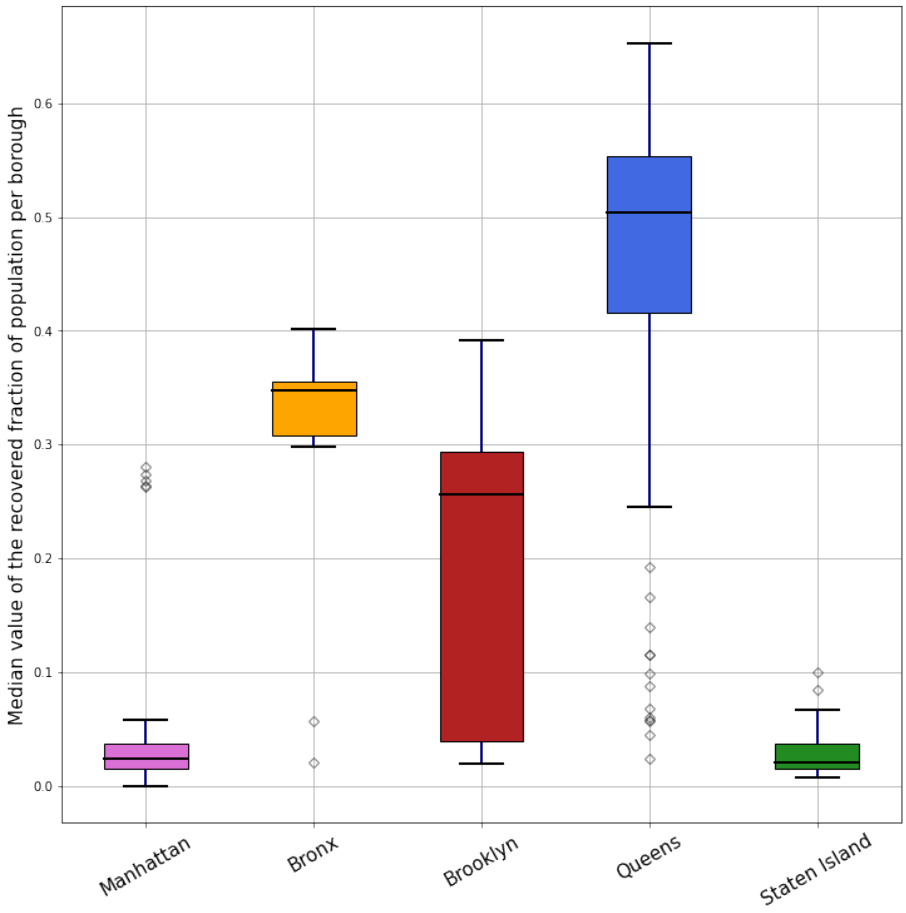

Figure 4.3: Statistical analysis of the simulation results in terms of fraction of recovered over the total population at borough level.

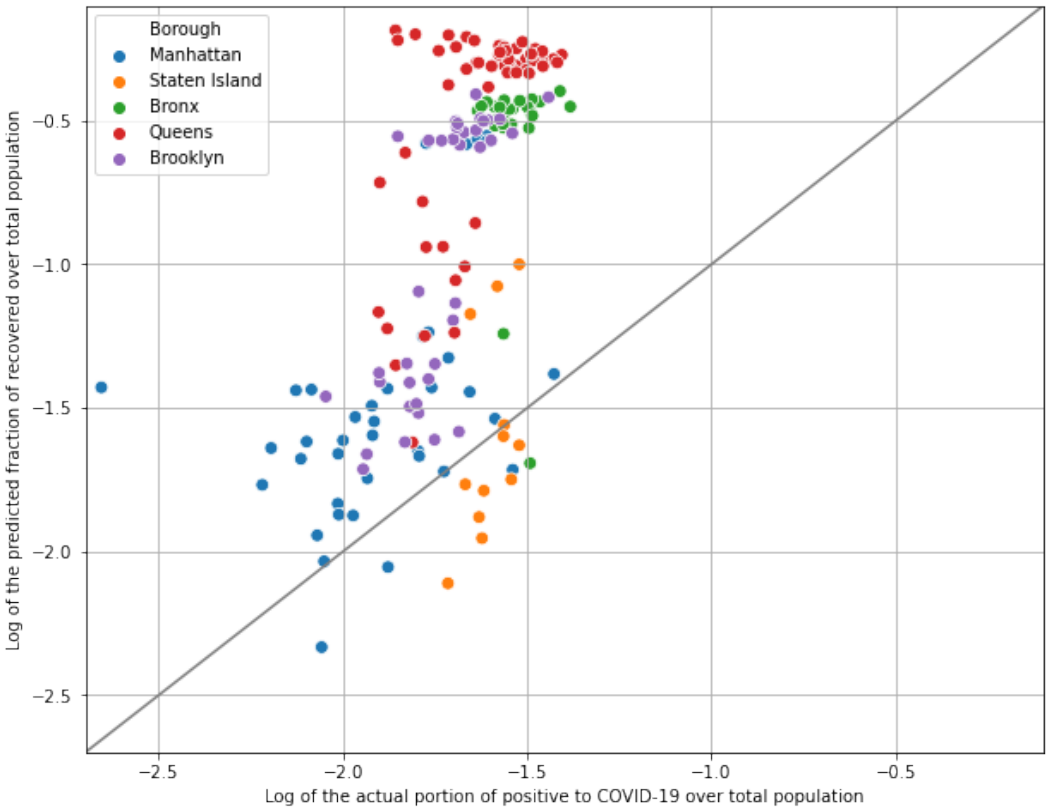

Figure 4.4: Scatter plot in logarithmic scale of predicted versus actual data of the portion of infected from COVID-19 over the total population per ZCTA.

been shown to be relevant for the spread of influenza at the national level [37] and accurate in the ILI epidemic dynamics modeling.
Secondly, the geographical association of a binary risk index (high or low) may have affected the results. In fact, by using a simple binary index, we are losing in granularity when comparing different ZCTAs. For example, almost all neighborhoods of Staten Island are classified as "low risk" areas. As result, the geographical differences in the infection incidence in this borough are less evident in map associated to the simulated data compared to the actual data.
Moreover, non - pharmaceutical interventions to prevent infection's spread cannot be summarized only by the variations over time of the commuting patterns. In fact, as underlined in the work [44], the increased awareness of the dangerous health consequences of COVID-19 may have contributed to further reductions in the effective transmission. In addition, the Public health safety guidelines implemented, such as the face-mask covering, have certainly contributed in preventing the possibility of direct transmission between individuals [45].
The transmissibility of a virus can be driven by environmental conditions, that depends on its characteristics. Currently a large number of studies are being carried out to deepen the specific environmental characteristics that have influenced the spread of the SARS-CoV-2, as the work [46].
Finally, the reaction diffusion process involved in the modeling, the discrete version of SIR model, is a very simple model and in general for COVID-19 dynamics it would be preferable to use more complex models, with an higher number of compartments, such as SEIR (susceptible, exposed, infected, recovered) models [47], [48].

Chapter 5

Summary and conclusions

In this study, we have proposed a mathematical model to predict the differences in COVID-19 incidence at spatial level for NYC neighborhoods between March and May 2020. First, we evaluated the correlations in the socio-economic differences at ZCTA level and the actual incidence of COVID-19. The multiple linear regression lead us to realise that commuting flows, median income and the possibility of social distancing associated to the different ethnic groups are determinant factors of the epidemic spread. This result is coherent with previous studies, [13], [23], [27] and supported the computational methods proposed in the second part of this work.

This study extends the existing literature on epidemiological modelling. A network structured meta-population model has been adapted to simulate the COVID-19 spread dynamics by adding the following features. First, the commuting flows between NYC ZCTAs are included in the model dynamics and updated at each week of simulation according to actual variations in mobility patterns. Secondly, we included in the model a binary risk index at ZCTA level to summarize the overall effects of the socio-economic disadvantage on the reaction-diffusion process. To sum up, the meta-population model qualitatively reproduce the actual geographical pattern of the COVID-19 spread across NYC neighborhoods.

The main limit of this study is represented by the overestimation in terms of numerical predictions of the cumulative incidence. This inaccuracy in predictions at spatial level could be overcome by introducing more than one source of data for the commuting patterns in the model. A pool of mobility variables could help in quantifying more accurately the actual commuting flows. Moreover, it could be useful to introduce a risk index with higher variability across ZCTAs, based on socio-economic characteristics. This might influence the reaction diffusion parameters and consequently the movements across the network. From an epidemiological point of view, this can surely help in forecasting

and controlling the future spread of high risk Influenza like-illness as COVID-19.

Bibliography

[1] *A Comparative Analysis of the Spanish Flu 1918 and COVID-19 Pandemics*, Akhilesh Agrawal at al. - The Open Public Health Journal, October 2020

[2] *Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China*, Joseph T. Wu, Kathy Leung, Mary Bushman et al. - Nature medicine, March 2020

[3] *Spatial epidemic dynamics of the COVID-19 outbreak in China*, Dayun Kang et al. - ScienceDirect, April 2020

[4] *Spatiotemporal analysis of COVID-19 outbreaks in Wuhan, China*, Wei Liu et al. - Nature, July 2021

[5] *Early dynamics of transmission and control of COVID-19: a mathematical modelling study*, Adam J Kucharski et al. - The Lancet, MArch 2020

[6] *The socio-spatial determinants of COVID-19 diffusion: the impact of globalisation, settlement characteristics and population*, Thomas Sigler et al. - BMC Globalization and Health, 2021

[7] https://coronavirus.health.ny.gov/new-york-state-pause

[8] *Spatial analysis of COVID-19 clusters and contextual factors in New York City*, J. Cordesa, M. C. Castro - NIH, August 2020

[9] https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

[10] https://www1.nyc.gov/site/doh/covid/covid-19-data-totals.page

[11] https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html

[12] *Global Progress on COVID-19 Serology-Based Testing*, Kobokovich A at al. - Johns Hopkins Center for Health Security, June 2020

[13] *Assessing capacity to social distance and neighborhood-level health disparities during the COVID-19 pandemic*, D. Carrión et al. - Medrχiv, June 2020

[14] *Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza*, C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller1, B. T. Grenfell - Science, April 2006

[15] *Detecting Robust Patterns in the Spread of Epidemics: A Case Study of Influenza in the United States and France*, P. Crépey, M. Barthélemy - American Journal of Epidemiology, October 2007

[16] *Multiscale mobility networks and the spatial spreading of infectious diseases*, D. Balcan, V. Colizza, B. Gonçalves et al. - PNAS, December 2009

[17] *Individual identity and movement networks for disease meta populations*, M. J. Keeling et al. - PNAS, December 2009

[18] *Modelling disease outbreaks in realistic urban social networks*, S. Eubank, H. Guclu et al. - Nature, 2014

[19] *Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis*, W. Yang, S. Kandula, M. Huynh, S. K Greene, G. Van Wye, W. Li, H. Tai Chan, E. McGibbon, A. Yeung, D. Olson, A. Fine, J.Sharman - The Lancet, February 2021.

[20] *Network structure and epidemic waves in meta population models*, V. Colizza, F. Gargiulo et al. - World Scientific, 2008/2009

[21] *Deciphering Dynamics of Epidemic Spread: The Case of Influenza Virus*, Ranjit Kumar Upadhyay et al. - World Scientific, 2014

[22] https://github.com/ClaudiaDeceglie/Thesis_NYC_covid19

[23] *Exposure density and neighborhood disparities in COVID-19 infection risk*, Boyeong Hong, Bartosz J. Bonczak, A. Gupta, L. E. Thorpe, and C. E. Kontokosta - March 30, 2021

[24] *Modelling and predicting the effect of social distancing and travel restrictions on COVID-19 spreading*, F. Parino, L. Zino, M. Porfiri and A. Rizzo - The Royal Society, February 2021

[25] https://opendata.cityofnewyork.us/

[26] https://www.census.gov/programs-surveys/acs

[27] *Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods*, George J. Borjas - April 2020

[28] *Disparities in COVID-19 Testing and Positivity in New York City*, W. Lieberman-Cribbin, S. Tuminello, R. M. Flores, E. Taioli - NCBI, June 2020

[29] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html

[30] *Analysing the link between public transport use and airborne transmission: mobility and contagion in the London underground*, L. Goscé, A. Johansson - December 2018.

[31] https://www.governor.ny.gov/sites/default/files/atoms/files/ PublicTransportation-MasterGuidance.pdf

[32] *U.S. county-level characteristics to inform equitable COVID-19 response*, T. Chin, R. Kahn, R. Li, J. T. Chen et al. - NIH, April 2020.

[33] https://www.nytimes.com/interactive/2020/05/15/upshot/who-left-new-york-coronavirus.html

[34] https://news.miami.edu/stories/2020/11/research-shows-covid-19-has-hit-hispanic-communities-hard.html

[35] *Modelling and predicting the effect of social distancing and travel restrictions on COVID-19 spreading*, F. Parino, L. Zino, M. Porfiri, A. Rizzo - The Royal Society, February 2021

[36] *Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models*, C. Panigutti, M. Tizzoni, P. Bajardi, Z. Smoreda, V. Colizza - The Royal Society, May 2017

[37] *Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza*, C. Viboud at al. - Science, April 2006

[38] *On the Use of Human Mobility Proxies for Modeling Epidemics*, M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, V. Colizza - PLOS Computational Biology, July 2014.

[39] *Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic*, Yuhao Kang, Song Gao et al.

[40] *Using Mobile Phone Data to Predict the Spatial Spread of Cholera*, L. Bengtsson, J. Gaudart, X. Lu, S. Moore et al. - Scientific reports, 2015.

[41] *The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology*, Kyra H. Grantz et al. - Nature, September 2020

[42] *Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis*, Miklos Szocska et al. - Nature, March 2021

[43] *Understanding individual human mobility patterns*, M.C. Gonzalez, C.A. Hidalgo, A.L. Barabasi - Nature, 2008.

[44] *Effectiveness of Non-pharmaceutical Interventions to Contain COVID-19: A Case Study of the 2020 Spring Pandemic Wave in New York City*, W. Yang, J. Shaff, J. Shaman

[45] *Use of facemasks during the COVID-19 pandemic*, H. J. Schünemann et al. - The Lancet, August 2020.

[46] *Environmental determinants of COVID-19 transmission across a wide climatic gradient in Chile*, F. Correa-Anareda et al. - Scientific reports, May 2021.

[47] *SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence*, A. Godio, F. Pace, A. Vergnano - IJERPH, May 2020

[48] *COVID-19 pandemic: a mobility-dependent SEIR model with undetected cases in Italy, Europe, and US*, N. Picchiotti, M. Salvioli, E. Zanardini, F. Missale - EP, September 2020