The heterogenous impact of Covid-19
Evidence from Italian municipalities

Francesco Armillei* Francesco Filippucci†

Preliminary version: August 3rd 2020

Abstract

The Covid-19 epidemic had an impact far from geographically homogeneous, even within most infected zones. We analyse the correlates of this heterogeneity at a very granular level, relying on a novel dataset with wide information on Italian municipalities. We first describe Covid-19 impact heterogeneities selecting a number of relevant covariates. We find that higher mortality rates across municipalities are associated with lower income, lower education, higher share of workers in industrial sector, lower household dimension, lower service and trade employment. This suggests that these areas are mostly peripheral ones. All our covariates are severely multicollinear, cautioning on causal interpretation of the results. As a second exercise, we use a machine learning methodology to predict areas with a high risk of Covid-related deaths independently from spatial proximity to infection. We believe that our findings might be useful to predict which areas are at higher risk given where the first outbreak occurs.

*Bocconi University

†Paris School of Economics

1 Introduction

It is well known that the spread of Covid-19 was concentrated in some regions within countries[1]. The reasons why Covid-19 spread dramatically in some regions and not in others is hard to evaluate and hard to identify: in Italy, for instance, the virus mostly spread in the regions of Lombardy, Emilia-Romagna and Piedmont. These regions are different from the rest of the country for several characteristics, for example income, weather, demographic profile. Also, although the first outbreak occurred in Codogno (Lombardy), in the middle of what became the most affected area, other outbreaks occurred for example in Veneto region, which reported a much lower death toll. It is nonetheless ascertained nowadays that the outbreak in Lombardy was underestimated and poorly managed. All these reasons suggest how difficult it is to draw conclusions about what favours the spread of Covid-19 basing on cross-regional comparison within a single country, an exercise that risks to be endogenous, low-power, and hardly credible.

In this article we do not focus on what causes variation across regions, but instead on the variation in Covid-19-related deaths and in relevant correlates *within* regions. We argue that this granular scale is a better one to gain insights on socio-economic environmental correlates of the spread of Covid-19. We highlight in fact that the increase of mortality due to Covid-19 varies significantly within highly infected regions, while a recent literature informed us of the importance of nighborhoods and granular data for a broad range of outcomes (Chetty et al. 2016).

We leverage a novel dataset assembled by the localopportunitieslab.it project, which is wide in the number of variables available at a municipality level. This allows us to have a comprehensive picture, in a similar spirit of Knittel and Ozaltun (2020) and Desmet and Wacziarg (2020) for the US. To the best of our knowledge, we are the first to run such an attempt in the Italian context. The Italian case is particularly interesting, being Italy the first Western

[1] https://www.europeandatajournalism.eu/eng/News/Data-news/A-fraction-of-European-regions-account-for-a-majority-of-COVID-19-deaths

country hit by the virus, so that in the first weeks following the discovery of the outbreak counter-measures were confused, late and arguably binding due to surprise. Such unfortunate situation for Italian population limits nonetheless the amount of confounding factors, such as policy responses, which threaten observational studies as ours.

As a first exercise, we explore how the impact of Covid-19 correlates with a number of selected variables of different socio-economic feature: income, employment, demographic structure, health facilities, pollution. We use a triple-diff and cross-sectional first-difference strategy. The results suggest that industrial employment share and average household dimension are positively associated with Covid-related deathsm while average income, education, service employment share, trade employment share, public transport index, house crowd index, average house price and maximum level of NO2 are negatively associated. The results seem to portray a picture where less developed and poorer municipalities are at higher risk.

Yet, we stress the fact that this first part of our inquiry is not causal. Actually, one of the main lessons is that one should be very cautious in considering socio-economic correlates of Covid-19, a popular exercise in last months. Because of the wide dataset we collect, we can show that there is severe multicollinearity between municipality-level variables such as average income, social capital, housing conditions, occupation, etcetera. We stress this since the number of studies considering correlates in isolation, or at most with a handful of controls, is surprising. Simple regressions can indeed represent insightful descriptives - and this is why we run them in the first part of our study – but one should keep them in prospective.

The second exercise – to the best of our knowledge so far unattempted in the literature – is instead one of predictive nature. We use our wide dataset to train a machine learning model including a large number of socio-economic environmental factors, plus a couple of spatial variables accounting for spatial correlation of extra mortality. We use Lasso shrinkage, and we either partial-out spatial variables before Lasso estimation, or we run the estimation only on highly-infected provinces. The selected non-spatial coefficients predict the risk of paying a high death toll to Covid19 on top of spatial proximity to highly-affected communes. This

methodology thus allows us to map on the whole Italian territory which municipalities are at higher risk if an outbreak occurs close to them.

Given the cost of lockdown policies, this is clearly a precious insight for policy-makers, who face a hard trade-off between the cost of Covid in terms of lives and socio-economic conditions. Selective lockdowns have by now been excluded as ineffective: even though locking down only individuals at high risk would reduce economic losses the virus would still spread regionally through asymptomatic or surviving infected. Our results suggest that we should consider the hypothesis of taking counter-measures basing on regional characteristics. This refers both to preventive action (testing, mask wearing, restrictions to gatherings) and to emergency measures such as lockdowns - perhaps basing on agglomerates of municipalities into local labor markets as proposed in Tortuga (2020).

2 Previous literature

This work is strongly related to the previous research in the economic field that has empirically highlighted potential links (i.e. correlations) between Covid-19 spread and several environmental characteristics. Summing up a burgeoning stream of literature, evidence emerged in favor of correlations between Covid-19 spread and a countless number of factors:

- **Income**: Brandily et al. (2020), using France municipal level data, find evidence of an income gradient in the impact of the pandemic on mortality: it is twice as large in the poorest municipalities compared to other municipalities. Borjas (2020) finds similar results studying the number of infections in the different neighbourhood of New York City. Knittel and Ozaltun (2020) find that US counties with higher home values have higher death rates.

- **Social capital**: Barscher et al. (2020), find that that high-social-capital areas exhibit lower excess mortality in a set of different European countries. Borgonovi et al. (2020), using data on US counties, find that disease spread was higher in high social capital areas, but mortality lower. Kuchler et al. (2020) relate positively Covid-19 spread and

social networks (measured with Facebook data) in Italy and in US.

- **Health facilities**:Sussman (2020), using cross-country evidence and controlling for a variety of contributing factors, finds that increasing the number of hospital beds has a significant and quite substantial impact on mortality rates. Alacevich et al. (2020) find that in Italy municipalities with care homes present significantly higher excess death rates among the elderly.

- **Urban structure**: Gerritse (2020), exploiting data on US counties, finds that population density is positively correlated to infection at the outbreak. Working on US counties as well, Knittel and Ozaltun (2020) find that higher amounts of commuting via public transportation, relative to telecommuting, is correlated with higher death rates. Desmet and Wacziarg (2020) find similar results. Sà (2020), leveraging on data from England and Wales, finds that contagion is higher where more people make use of public transportation. Carozzi et al, (2020) find that density has affected the timing of the outbreak in each county, with denser locations more likely to have an early outbreak, but do not find any impact on COVID-19 cases and deaths.

- **Demographic profile**: Knittel and Ozaltun (2020), looking at US counties, find that a higher presence of elderly people positively correlates with death rate. Desmet and Wacziarg (2020) finds similar results. Borjas (2020) finds in New York City a similar correlation between Covid-19 cases and the presence of larger households. Sà (2020) highlights similar correlations in England and Wales. Aparicio and Grossbard (2020) find for a series of European countries and most of all for the US that more people died from Covid in countries or states with higher rates of intergenerational co-residence.

- **Pollution**: Cole et al. (2020) find positive relationship between Covid-19 deaths and cases and air quality in the Netherlands, particularly concentration of PM2.5. In Germany, Isphording and Pestel (2020) find significant positive effects of PM10 concentration after the onset of the illness on COVID19 deaths specifically for elderly patients (80+ years).

- **Etnicity**: Borjas (2020) finds that people living in in neighbourhoods of New York

City with predominantly black populations were less tested and more likely positive. Knittel and Ozaltun (2020) find that higher shares of African American residents in the county are correlated with higher death rates.

- **Weather**: Kapoor et al. (2020) find that rainfall impacts positively on social distancing and causes a reduction in cases and deaths that persists for weeks. Knittel and Ozaltun (2020) find that US counties with higher summer temperatures and lower winter temperatures have higher death rates.

This stream of literature seems still to be proceeding by trials and errors. With some exceptions, all these studies lack identification and thus should not be interpret as causal. Moreover, few of them control for variables other than average age. Thus, despite the fact that literature is nowadays crowded, it lacks a comprehensive picture and it is often unclear about the value of lessons learned.

3 The Italian context

The first Covid case in Italy was officially detected on the 30th of January 2020, at the Spallanzani Institute of Rome, after the couple travelled from Wuhan to Milan, Verona, Parma and Florence. On the same day, the Italian government enacted a ban for all flights arriving from China, whose duration was set to 90 days. The first case of secondary transmission was detected on the 18th of February in Codogno, a municipality in the region of Lombardy. As a response, the Italian Government imposed first local lockdowns in those municipalities hit by the Covid-19. However, the virus continued spreading, especially in the northern part of the country. The 8th of March the entire region of Lombardy (which alone accounts for roughly the 17% of population of Italy and the 22% of GDP) was locked down. Two weeks later, on the 22nd of March, the Italian government took the unprecedented decision to prohibit all individual on Italian soil from travelling in a municipality different that where they were, a part for working or health reasons. Besides, every non-necessary economic activity was shut down. This marked the beginning of the lockdown of the entire country and of the so called "Phase 1" of the epidemy.

4 Data

For this study we rely on two main data sources. First, the official dataset provided by the Italian National Statistical Institute (ISTAT) containing the daily count of death for each Italian municipality. This dataset allows us to perform a very granular analysis of the period in which the Covid-19 crisis hit the country. For each municipality we observe the daily number of deaths (however, without knowing the cause), by gender and age, for multiple years (despite we focus only on 2019 and 2020). One should keep in mind that every death is registered in the municipality of residence of the dead.

The second data source is the Local Opportunities Lab (from now on LOL) dataset. This newly available dataset gathers information from a number of public sources at a municipality level (mainly census, fiscal data, or official statistics by Istat), with information that ranges from housing to education to income. Data are cleaned and re-elaborated by the Local Opportunities Lab (for example, dealing with municipalities that have merged or split). The LOL dataset represents a useful step forward in order to study socio-economic phenomena in Italy and this work represents the first attempt to fully exploit its potential. For more information on the Local Opportunities Lab please visit `https://www.localopportunitieslab.it/`.

We merge the two datasets thanks to uniquely identifying municipality ID, dropping municipalities that miss data. In the merged dataset we observe 7172 municipalities (out of 7914), which cover 94% of the Italian population. Our analysis of Covid impact focuses on 2020: however, we do not have data for 2020 in the LOL dataset. We then decide to associate to each municipality the most recent value of every covariates, assuming that any possible variation since then would not alter our results. Table 1 provides summary statistics of the correlates we use in our analysis.

We do not observe all variables for all municipalities. In particular, we have data only for a subset of the municipalities in our dataset when it comes to pollution, because the

Table 1: Descriptive statistics

VarName	Obs	Mean	SD	Min	Median	Max
Avg_income	7151	17083.99	3662.701	6213.882	17213.48	47807.83
Share_60_70	7129	0.13	0.019	0.026125	0.128543	0.261147
Share_70_80	7130	0.1	0.022	0.016691	0.102428	0.268817
Share_over80	7127	0.08	0.029	0.018385	0.077831	0.287671
Share_old	7118	0.25	0.052	0.058781	0.24468	0.499253
Avg_househ._dim.	7162	2.36	0.262	1.2	2.4	3.4
Education	7162	136.97	49.592	28.4	129.7	568.5
Mobility	7162	59.89	8.447	13.5	61.6	79.4
Outer_mobility	7162	35.09	12.591	0	36.2	67.5
Public_mobility	7162	11.44	4.658	0	10.6	50.9
Mobility_slow	7162	16.84	7.931	0	15.6	75.2
Drug_stores	7162	0.5	0.577	0	0.3	10.3
Gini	7147	0.19	0.02	0.1262	0.1891	0.3141
Unemp.t_share	7162	10.17	6.311	0	7.8	42.2
Employment_share	7162	45.16	7.883	18	46.5	70.1
Agri._empl.	7162	9.12	8.474	0	6.325	78.5
Industrial_empl.	7162	31.25	10.805	5.5	30.9	75
Trade_employment	7162	18.77	5.083	1.6	18.1	68.5
Service_employment	7162	40.86	8.807	9.5	40.1	75.8
Density_index	7162	307.07	653.16	1	109.05	12224.4
House_crowd_index	7162	0.42	0.423	0	0.3	6
Beds_acute_area	5295	0	0	0	0.000204	0.00331
Beds_acute	5295	2122.3	6127.246	0	1086	234867
Beds_acute_average	5295	0.89	8.094	0	0.333333	579.9185
Beds_area	5295	0	0	4.34E-06	0.000297	0.003586
Beds_total	5295	3094.48	8815.562	2	1538	331047
Beds_average	5295	1.26	11.398	0.000309	0.480329	817.4
Avg_house_price	7157	121467.9	69323.81	15832.13	110473.5	1309432
Mean_BAP	1346	0.55	0.399	0	0.4	2.2
Mean_O3	2471	53.02	10.09	33.43471	50.77158	97.63878
Mean_PM10	2773	25.74	6.464	5	26	51
Mean_PM25	2011	17.4	4.533	5	18	27
Max_PM10	2773	92.13	32.958	21	90.67632	307.6
Max_NO2	2926	109.35	30.703	6.57824	110.9925	212.5
Cultural_empl.	7170	0.02	0.05	0	0.00514	0.839806
Association_employment	7170	0.12	0.127	0	0.079696	0.930435
Volunteer_share	7170	0.27	0.168	0	0.233931	0.930435

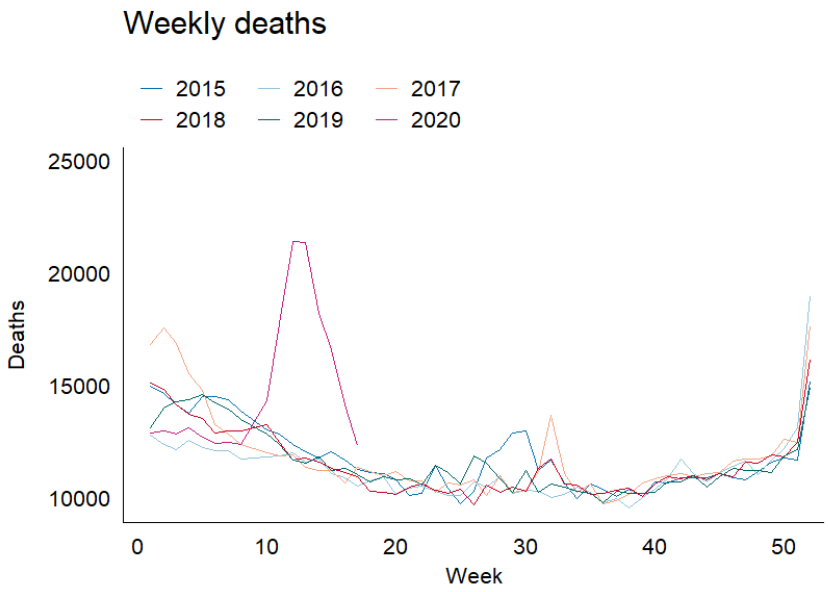
measurement stations are located only in some municipalities and we drop municipalities further than 30 mins of car from a measurement station. Hospital beds are also considered

only in the range of 30 minutes from the municipality.

5 Empirical strategy

Concerning the period of analysis, we decide to focus our attention on the month of March. In Italy, this corresponded to the peak of the spread of Covid-19, as highlighted by Figure 1. Moreover, deaths during this period are likely resulting from a contagion precedent to the national lockdown, which could act as confounding, undermining our analysis of the factors correlating with the spread of Covid-19. In fact, there is growing evidence that the citizens' response to lockdown and the effect of such policies have been far from homogeneous across different zones (for the Italian case see for example Di Porto et al., 2020) and correlated with several socio-economic features.

Figure 1



A key feature of our analysis is how we measure our outcome of interest: the harshness of Covid-19 epidemic in a municipality. We decide to use the rate of death (number of deaths divided by total population) in March and to see how this value varies in 2020 with respect to the average of 2017-19. From now on, we will refer to this difference as

extra-mortality. This measure, of course, can be noisy: in particular, although using death rates rather than death count allows to have comparable average levels of the outcome across municipalities of different sizes, the outcome can be much more volatile for smaller municipalities. Also, although minimizing deaths can be considered the chief policy goal, there is increasing evidence that even non-deadly contagion causes severe health damages, raising the doubt that one should consider contagions as an outcome rather than deaths. However, contagion measurements are also noisy, partially endogenous, and harder to obtain. Moreover, the Italian government released data on the number of Covid-19-related deaths aggregated only at a regional level (NUTS-2) and data on the number of Covid-19 cases at provincial level (NUTS-3). For the scope of our work, these two levels of analysis would be too aggregated. It is also worthwhile noticing that official data on Covid-19 infections and death may be imprecise and endogenous, since the screening systems covers a small fraction of the population, particularly in the early phase of the epidemic, and such fraction is highly dependent on regional health policy. To be sure, we compare our measure of extra-mortality and aggregated data on contagion in the month of March, finding that our measure correlates well with the data on regional deaths and provincial infection cases. Figure 2 and 3 provide graphical evidence.

Figure 2

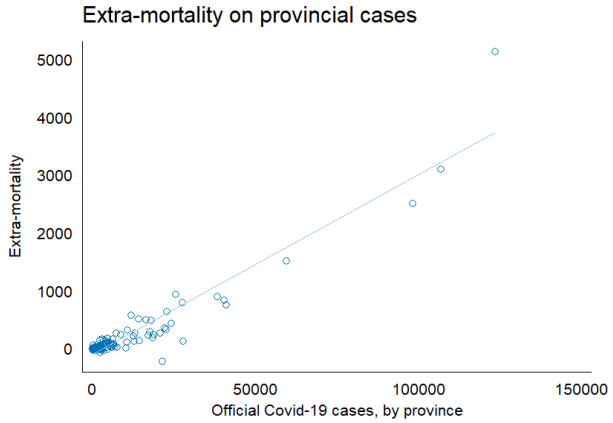
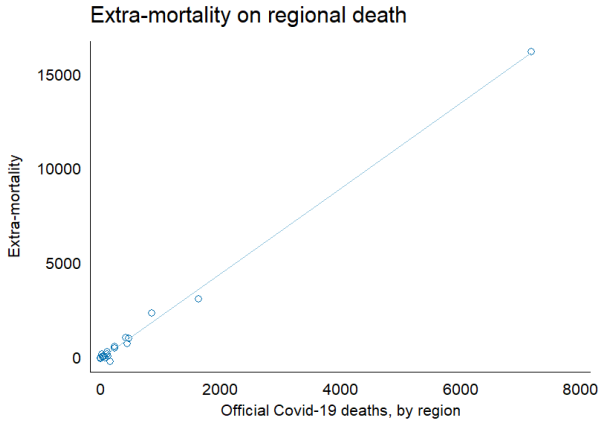


Figure 3



Turning to the geographical side of the contagion, Figure 4 reports the level of extra-mortality in March 2020 relative to an average of March 2017-19 by municipality across Italy. In the

early phase of the epidemic, the virus spread mostly in the north of the country (especially in Lombardy) but we cannot distinguish whether this is endogenous or due to a first exogenous outbreak in that area. For this reason, comparing a municipality in a high-contagion province with one in a province where the virus arrived later can be misleading. To address this issue, we distinguish high-infection provinces and low-infection provinces, and exploit variation within high-infection provinces. High infection provinces are defined as those where mortality at a provincial level increases more than the 75th percentile and correspond mostly to Lombardy, parts of Emilia and Piedmont, and the province of Ancona in Central Italy. Figure 5 pictures the extra-mortality in these provinces. It is noteworthy that in low-infection provinces mortality is almost unchanged.

Figure 4

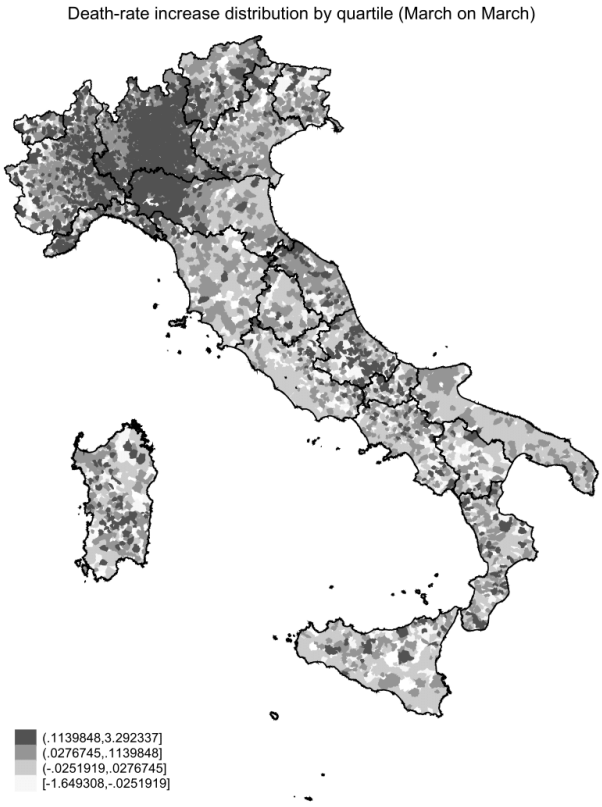
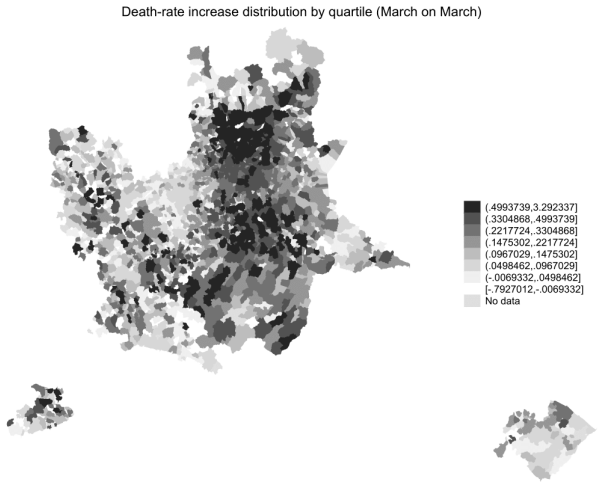


Figure 5



5.1 Triple differences approach

In order to see what environmental features of Italian municipalities correlate with the number of Covid-related deaths, we first adopt a triple-differences approach, in the spirit of Brandily et al. (2020) using municipalities as unit of analysis. This approach, despite being simplistic, has the privilege of being neat.

Firstly, we define as "post" the death rate in March 2020 and as "pre" the average death rate in the same months in 2017-2019. Then, since we want to assess the effect on extra mortality due to the epidemic, our second level of differentiation is between high and low infection provinces. This also allows us to control for possible shocks affecting both high and low infection provinces at the time of Coronavirus such as the lockdown. Thirdly, to understand the role of each of our correlates of interest, we divide our observations in two groups according to the level of the correlate. We classify a municipality as "high type" if its value of the correlate is above the 75th percentile, and "low type" if it is below. Given these groups, we compare the death-rate per 1000 of inhabitants pre and post Covid-19 (hence in 2017-2019 average and in 2020) in high versus low infected municipalities and in high versus low type municipality. The underlying assumption of this approach is that the difference in the evolution of deaths in March between high and low type municipality would have been the same in high and low infected provinces in the absence of the Covid-19 shock.

Given that we believe that at this point of research on Covid-19 it is possible to find only (hopefully useful) correlations, we adopt an "agnostic approach" and we analyse our correlates one by one, without employing controls. Hence, for each correlate of interest we run the following regression:

$$DeathRate = \alpha + \beta_1 Y + \beta_2 I + \beta_3 C + \beta_4 (Y * I) + \beta_5 (Y * C) + \beta_6 (I * C) + \beta_7 (Y * I * C) + \delta FE + \epsilon$$

where Y is a year dummy, I is an high/low infection provinces dummy, C is dummy for high/low type municipalities and FE are regional fixed effects. Our coefficient of interest is β_7, which captures the effect of our correlate of interest. We weight the regression by the population of each municipality. We cluster the standard errors at provincial level. Table 2 sums up the result of the coefficient of interest for such a regression of each of the correlates.

The correlates which turned out to be significant at a level of .05 are the following: average household dimension (positive coefficient), public transport index, education, service employment share, average income, house crowd index (negative coefficients). Including those significant at a level of .1 we add: industrial employment share and cultural employment (positive coefficients) and average house price, maximum level of NO2 and trade employment share (negative coefficients). On the whole, these results seem to suggest that richer, more educated and service-oriented areas were less hit by the Covid-19.

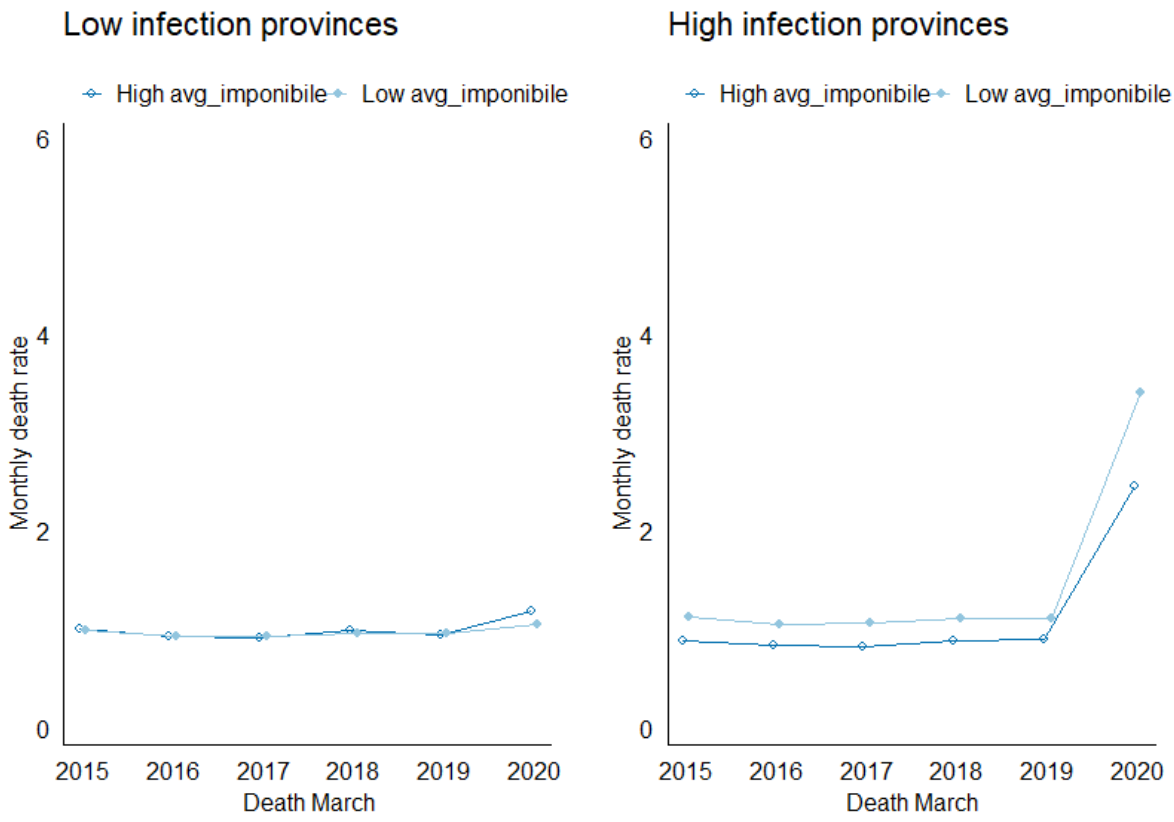
Using the average income variable as an example, the intuition for interpreting each coefficient produced by our triple-diff approach is straightforward by looking at Figure 6. On the left of such figure one can see the average mortality in March for municipalities belonging to low infection provinces, both for municipalities with high and low average income, over the years 2015-2020. We can see how there is only a small change in mortality rates in 2020, slightly stronger for high income municipalities. On the right exhibit we see the same in high infection provinces: the increase for low-income municipalities is larger than for high-income ones, determining the significance of our coefficient in Table 2. It should be noted that in high-infection areas mortality in low-income municipalities was already higher, but in a

Table 2: Triple difference approach - Coefficients of interest

Correlate	coef	stderr	pval
Avg_household_dimension	0.843	0.308	0.00727
Public_mobility	-0.862	0.317	0.00762
Education	-0.984	0.370	0.00896
Service_employment	-0.866	0.361	0.0181
House_crowd_index	-0.655	0.279	0.0209
Avg_income	-0.885	0.422	0.0384
Industrial_employment	0.897	0.471	0.0596
Trade_employment	-0.403	0.212	0.0605
Avg_house_price	-0.953	0.529	0.0742
Max_NO2	-0.455	0.270	0.0947
Cultural_employment	0.481	0.290	0.100
Mean_O3	0.533	0.333	0.112
Mobility	-0.617	0.389	0.115
Gini	-0.293	0.186	0.118
Outer_mobility	-0.339	0.223	0.132
Volunteer_share	0.594	0.391	0.132
Unemployment_share	-0.709	0.493	0.153
Max_PM10	0.496	0.350	0.159
Density_index	-0.527	0.373	0.160
Drug_stores	0.424	0.328	0.199
Share_60_70	0.422	0.359	0.242
Beds_total	-0.551	0.496	0.270
Employment_share	-0.279	0.256	0.278
Beds_acute	-0.524	0.491	0.289
Agricoltural_employment	0.539	0.506	0.289
Share_70_80	-0.201	0.202	0.323
Beds_area	-0.578	0.584	0.324
Beds_acute_area	-0.545	0.577	0.347
Association_employment	0.325	0.345	0.348
Mean_PM10	-0.340	0.424	0.424
Mean_BAP	0.307	0.415	0.461
Beds_average	-0.224	0.320	0.485
Beds_acute_average	-0.249	0.356	0.486
Mean_PM25	-0.102	0.291	0.728
Share_old	-0.109	0.362	0.763
Mobility_slow	0.0614	0.262	0.816
Share_over80	-0.0121	0.444	0.978

clearly parallel way. This further justifies our triple-diff approach.

Figure 6



5.2 Cross-sectional approach

As an alternative approach, we proceed with a cross-sectional regression of first differences of our outcome on the same selected covariates of the previous section. In this case we define as our dependent variable the extra-mortality rate at municipal level, computed as the difference between the mortality rate in March 2020 and the average in March 2017-2019 (and then multiplied by 1000). We restrict our sample only to municipalities in high-infected provinces. Similarly to before, we add regional fixed effect, we cluster standard error at

provincial level, we weight the regression by the population of the municipality and we run a separate regression for each one of our correlate. However, differently from before, we do not "discretize" our correlates of interest (i.e. with do not split the municipality in high and low type). Table 3 sums up the coefficient of each correlate of interest.

The correlates which turned out to be significant at a level of .05 and show a negative coefficient are: average income, education, service employment share, employment share, Gini index, beds in acute care, total hospital beds, urban density index, public transport index, mobility index, house crowd index, average house price, trade employment share. The correlates which turned out to be significant at a level of .05 and show a positive coefficient are: share of citizens between 60 and 70 years old, industrial employment share, average household dimension, volunteer share, association employment, drug stores per 1000 inhabitants. Including those significant at a level of .1 we add: agricultural occupation share (positive coefficient) and maximum level of NO2 (negative coefficient).

On the whole, these results seem consistent with those obtained through the triple-differences approach. This reassures us, as the results do not seem to strongly depend on the discretization of the outcome or on changes in the low-infection areas. To sum up, the correlates that turned out to be statistically significant in both analyses are: industrial employment share, average household dimension, (positive coefficient) and average income, education, service employment share, trade employment share, public transport index, house crowd index, average house price and maximum level of NO2 (negative coefficients).

Our results seem to confirm those of Brandily et al. (2020) on the role of income while we find a negative association between Covid-related deaths and house prices, differently from Knittel and Ozaltun (2020). We believe that coefficients of the same sign between average income and average house price are consistent. When it comes to urban structure, our results differ significantly from those obtained by Gerritse (2020), Knittel and Ozaltun (2020) Desmet and Wacziarg (2020) and Sà (2020), as we find that public transport and crow are negatively correlated with Covid-related deaths. These differences might stem firstly from

Table 3: Cross-sectional approach - Coefficients of interest

Correlate	coef	stderr	pval
Avg_income	-0.000140	3.95e-05	0.00232
Education	-0.00687	0.00199	0.00286
Service_employment	-0.0507	0.0149	0.00309
Employment_share	-0.0996	0.0300	0.00375
Gini	-12.55	3.911	0.00485
Beds_acute	-0.000173	5.28e-05	0.00503
Share_60_70	30.95	9.721	0.00514
Beds_total	-9.62e-05	3.00e-05	0.00588
Density_index	-0.000283	9.27e-05	0.00685
Industrial_employment	0.0537	0.0183	0.00890
Public_mobility	-0.0669	0.0232	0.00977
Mobility	-0.139	0.0503	0.0127
House_crowd_index	-1.626	0.589	0.0129
Avg_house_price	-9.82e-06	3.60e-06	0.0138
Volunteer_share	5.508	2.278	0.0264
Avg_household_dimension	1.642	0.695	0.0296
Association_employment	4.334	1.870	0.0324
Trade_employment	-0.0537	0.0252	0.0467
Drug_stores	1.097	0.521	0.0495
Max_NO2	-0.0148	0.00727	0.0575
Agricoltural_employment	0.140	0.0714	0.0652
Unemployment_share	-0.143	0.0876	0.119
Share_old	4.931	3.330	0.156
Outer_mobility	0.0109	0.00875	0.229
Beds_acute_average	0.142	0.121	0.257
Mean_BAP	-0.822	0.731	0.280
Mean_O3	0.0498	0.0478	0.312
Beds_average	0.0462	0.0515	0.384
Cultural_employment	2.425	2.899	0.414
Share_70_80	6.114	7.334	0.415
Mobility_slow	0.0179	0.0221	0.429
Share_over80	3.877	8.395	0.650
Beds_area	483.3	1278	0.711
Mean_PM10	-0.0177	0.0530	0.742
Beds_acute_area	400.6	1242	0.751
Max_PM10	0.000746	0.0161	0.964
Mean_PM25	0.000513	0.0626	0.994

our granular approach (while other works rely on county data) and secondly from our "agnostic" approach (we test correlates one by one, without using controls). In fact, income and higher density are collinear and hard to disentangle. With respect to the previous literature, we highlight that the sectoral composition of the economic system matters.

As an important concluding remark to our empirical analysis, Table 4 shows the correlation coefficients between our correlates of interest. The color of each cell refers to the statistical significance of the correlation (from red to green).It clearly emerges that most of our variables are significantly correlated. At this stage of the research, it is quite hard to disentangle the effect on Covid-19 diffusion of each of them with simple regressions and we argue that many of the studies in the literature we reviewed underestimate this factor.

Table 4: Correlation matrix

	Avg_income	sh._60_70	sh._70_80	sh._over80	sh._old	Avg_hou.	Education	Mob.	Outer_Mob.	Public_Mob.	Mob._slow	Drug_stores	Gini
Avg_income	1												
sh._60_70	0.0387	1											
sh._70_80	0.369	0.536	1										
sh._over80	0.135	0.573	0.739	1									
sh._old	0.245	0.699	0.913	0.925	1								
Avg_hous.n	-0.473	-0.416	-0.65	-0.709	-0.714	1							
Education	0.432	0.136	0.332	0.274	0.3	-0.325	1						
Mob.	0.769	-0.02	0.265	0.133	0.189	-0.491	0.102	1					
Outer_Mob.	0.387	0.101	0.0956	-0.012	0.0626	-0.214	-0.188	0.581	1				
Public_Mob.	-0.01	-0.025	0.0289	0.0272	0.0122	0.0147	0.204	-0.162	-0.157	1			
Mob._slow	-0.074	-0.196	-0.11	-0.095	-0.135	0.199	-0.082	-0.233	-0.516	0.145	1		
Drug_stores	-0.206	0.282	0.128	0.331	0.259	-0.217	-0.003	-0.149	0.0889	0.0193	-0.085	1	
Gini	0.109	-0.142	-0.079	-0.113	-0.13	0.169	0.501	-0.268	-0.431	0.246	0.181	-0.156	1
Unempl._sh.	-0.663	-0.192	-0.411	-0.33	-0.39	0.643	-0.086	-0.871	-0.538	0.234	0.289	0.0487	0.303
empl._sh.	0.68	-0.052	0.0927	-0.012	0.0351	-0.427	-0.004	0.93	0.603	-0.258	-0.271	-0.129	-0.302
Agric._empl.	-0.548	0.0267	-0.181	0.0812	-0.044	0.158	-0.253	-0.385	-0.237	-0.066	0.0982	0.217	-0.07
Industrial_empl.	0.249	-0.035	-0.008	-0.105	-0.047	-0.021	-0.486	0.467	0.459	-0.336	-0.118	-0.122	-0.466
Trade_empl.	-0.021	-0.032	0.0479	-0.02	0.0011	-0.047	0.039	0.0145	0.0006	0.0733	-0.077	0.0045	0.141
Service_empl.	0.117	0.0314	0.119	0.067	0.0824	-0.071	0.7	-0.251	-0.342	0.391	0.0901	-0.019	0.513
Density_index	0.0733	-0.253	-0.085	-0.242	-0.201	0.304	0.0674	-0.175	-0.198	0.353	0.447	-0.189	0.225
House_crowd_index	-0.331	-0.236	-0.32	-0.321	-0.348	0.507	-0.181	-0.528	-0.379	0.21	0.466	0.0085	0.181
Beds_ac._area	-0.147	0.276	0.202	0.318	0.29	-0.328	0.108	-0.05	-0.107	-0.166	-0.167	0.166	-0.079
Beds_ac.	0.419	-0.261	-0.032	-0.226	-0.169	0.124	0.0662	0.257	0.134	0.323	0.176	-0.269	0.0788
Beds_ac._av.	0.163	0.149	0.0274	0.0794	0.0757	-0.083	-0.096	0.206	0.471	-0.093	-0.157	0.152	-0.258
Beds_area	-0.179	0.283	0.201	0.337	0.3	-0.331	0.123	-0.077	-0.12	-0.151	-0.194	0.192	-0.078
Beds_total	0.386	-0.276	-0.061	-0.25	-0.197	0.153	0.0623	0.193	0.11	0.379	0.23	-0.255	0.0945
Beds_av.	0.168	0.113	-0.002	0.045	0.0403	-0.057	-0.103	0.189	0.474	-0.056	-0.125	0.155	-0.258
Avg_house_price	0.638	-0.073	0.267	0.0524	0.131	-0.24	0.35	0.488	0.0923	0.196	0.0648	-0.26	0.311
*_BAP	0.0004	0.0227	0.033	-0.031	0.0067	-0.052	-0.001	0.0757	0.0954	0.0497	-0.197	-0.066	-0.067
*_O3	-0.378	-0.018	-0.052	0.0328	-0.017	0.128	0.0231	-0.443	-0.321	0.0679	-0.06	0.0811	0.263
*_PM10	0.0375	-0.279	-0.22	-0.373	-0.331	0.408	-0.162	-0.047	0.0462	0.136	0.253	-0.187	0.0052
*_PM25	0.397	-0.088	0.0844	-0.087	-0.009	-0.026	-0.19	0.465	0.352	-0.094	0.0375	-0.17	-0.225
Max_PM10	-0.289	-0.331	-0.39	-0.464	-0.466	0.619	-0.173	-0.462	-0.232	0.103	0.314	-0.14	0.184
Max_NO2	0.243	-0.15	-0.14	-0.226	-0.202	0.145	0.0954	0.102	0.0996	0.26	0.189	-0.157	0.178
Cultural_empl.	-0.062	0.0809	0.0472	0.177	0.123	-0.092	0.0815	-0.054	-0.11	0.0844	0.0433	0.127	0.0946
Ass._empl.	-0.031	0.192	0.172	0.248	0.233	-0.212	0.0365	0.0505	0.127	-0.075	-0.152	0.209	-0.099
Volun._sh.	0.012	0.256	0.226	0.313	0.299	-0.339	0.067	0.119	0.184	-0.085	-0.207	0.164	-0.106

	Unempl._sh.	empl._sh.	Agri._empl.	Industrial_empl.	Trade_empl.	Service_empl.	Density_index	House_crowd*	Beds_ac._area	Beds_ac.	Beds_ac._av.	Beds_area
Unempl._sh.	1											
empl._sh.	-0.877	1										
Agricoltural_empl.	0.267	-0.314	1									
Industrial_empl.	-0.462	0.523	-0.358	1								
Trade_empl.	-0.001	0.0334	-0.203	-0.232	1							
Service_empl.	0.324	-0.37	-0.231	-0.77	0.038	1						
Density_index	0.356	-0.271	-0.218	-0.161	0.0228	0.322	1					
House_crowd_index	0.6	-0.512	0.157	-0.212	-0.047	0.141	0.507	1				
Beds_ac._area	-0.13	0.005	0.168	-0.127	0.0764	-0.006	-0.306	-0.145	1			
Beds_ac.	-0.04	0.175	-0.367	0.122	0.0211	0.114	0.544	0.191	-0.37	1		
Beds_ac._av.	-0.207	0.221	0.0635	0.192	-0.145	-0.204	-0.148	-0.124	-0.012	0.0809	1	
Beds_area	-0.106	-0.028	0.194	-0.162	0.0725	0.017	-0.33	-0.154	0.99	-0.416	-0.017	1
Beds_total	0.0219	0.108	-0.342	0.0731	0.0019	0.158	0.602	0.243	-0.405	0.965	0.0492	-0.445
Beds_av.	-0.182	0.205	0.0637	0.192	-0.164	-0.197	-0.121	-0.104	-0.077	0.113	0.982	-0.077
Avg_house_price	-0.387	0.404	-0.401	0.0069	0.267	0.176	0.193	-0.102	-0.107	0.379	-0.037	-0.134
*_BAP	-0.139	0.125	-0.087	0.136	-0.032	-0.078	-0.136	-0.215	0.166	0.0295	-0.037	0.142
*_O3	0.397	-0.465	0.215	-0.349	0.102	0.197	0.0301	0.211	0.121	-0.211	-0.126	0.153
*_PM10	0.162	-0.064	-0.053	0.135	-0.185	-0.046	0.273	0.164	-0.482	0.383	0.0473	-0.508
*_PM25	-0.397	0.456	-0.16	0.495	-0.129	-0.389	0.0499	-0.224	-0.357	0.383	0.152	-0.41
Max_PM10	0.57	-0.459	0.0456	-0.117	-0.122	0.141	0.437	0.474	-0.305	0.309	-0.071	-0.319
Max_NO2	0.0726	0.0528	-0.318	0.0349	0.0496	0.167	0.331	0.134	-0.372	0.392	0.0625	-0.391
Cultural_empl.	0.0031	-0.069	0.0319	-0.125	0.0064	0.114	-0.063	0.0243	0.122	-0.094	-0.051	0.14
Ass._empl.	-0.137	0.0637	0.0842	-0.003	-0.068	-0.031	-0.224	-0.138	0.251	-0.2	0.14	0.25
Volun._sh.	-0.239	0.13	0.0865	0.0095	-0.059	-0.05	-0.304	-0.207	0.295	-0.27	0.0989	0.299

	Beds_total	Beds_av.	Avg_house_price	*_BAP	*_O3	*_PM10	*_PM25	Max_PM10	Max_NO2	Cultural_empl.	Ass._empl.	Volunteer_sh.
Beds_total	1											
Beds_av.	0.108	1										
Avg_house_price	0.347	-0.048	1									
*_BAP	0.0008	-0.042	0.0347	1								
*_O3	-0.19	-0.142	-0.139	-0.272	1							
*_PM10	0.407	0.108	-0.015	0.222	-0.323	1						
*_PM25	0.349	0.174	0.156	0.272	-0.477	0.619	1					
Max_PM10	0.328	-0.036	-0.148	0.0554	0.024	0.697	0.12	1				
Max_NO2	0.416	0.0983	0.265	-0.04	-0.27	0.363	0.198	0.253	1			
Cultural_empl.	-0.099	-0.055	0.0254	0.0012	0.0138	-0.156	-0.149	-0.089	-0.021	1		
Assoc._empl.	-0.222	0.104	-0.018	0.0009	0.0021	-0.222	-0.134	-0.222	-0.138	0.0683	1	
Volunt._sh.	-0.294	0.0663	0.0355	0.0353	-0.02	-0.329	-0.16	-0.329	-0.184	0.246	0.741	1

6 A machine learning exercise

As a second exercise, we develop a methodology to predict the risk of Covid-19 related deaths at municipality level, independently from where the first contagion occurs. This means that one can identify which municipalities risk paying a higher death toll in general and should thus be wearier if an outbreak occurs nearby. Our methodology goes as follows:

1. Using our wide dataset, we train a machine learning model of prediction of higher increase in the death rates, including spatial controls for the severeness of mortality increase in the surroundings (using driving time to define them).

2. Then, from the selected model, we drop spatial controls, and calculate predicted values using the remaining variables. This yields the level of risk given spatial proximity to other affected communes.

The kitchen sink of our machine learning exercise includes the following variables as non-spatial correlates:

- Per-capita income from work, housing, pensions, independent activity, entrepreneurial activity;

- Various measurements of housing crowding, household size and composition, residents' occupation and education;

- Demographic information: percentage of male, percentage of people in each 5-years age bracket;

- Pollution: mean levels of BAP, O3, PM10, PM2.5, numbers of days above limits of PM10 and NO2;

- Hospital beds per resident in the municipality, average number of hospital beds in 30 minutes driving range, per individual in 30 minutes range from the hospital, for normal and acute therapy, and from public and private structures;

- We also include squares of all variables and quintiles of total income, education, number of commuters out of the municipality, number of users of public transportation.

There are clearly several degrees of freedom in deciding which socio-economic variables to include, which machine learning method to use and which kinds of spatial correlates to include. Concerning the shrinkage algorithm, we chose to use a rigorous lasso, which selects a data-driven penalization and allows for heteroskedastic and clustered errors (Belloni et al., 2012). This method is easy to implement, relatively well-studied, and allows weights for population. Concerning the kind of spatial correlates in the first step of our methodology we run two attempts:

1. A regularized lasso of extra-mortality on the set of correlates, using all areas but partialling out two spatial controls: the number of deaths in municipalities reachable by car in 15 and 30 minutes.

2. A regularized lasso of extra-mortality on the set of correlates and two spatial controls (number of deaths in municipalities in the ray of 15 and 30 minutes), using only high-infection areas.

Table 5 shows the selected coefficient using the first method.

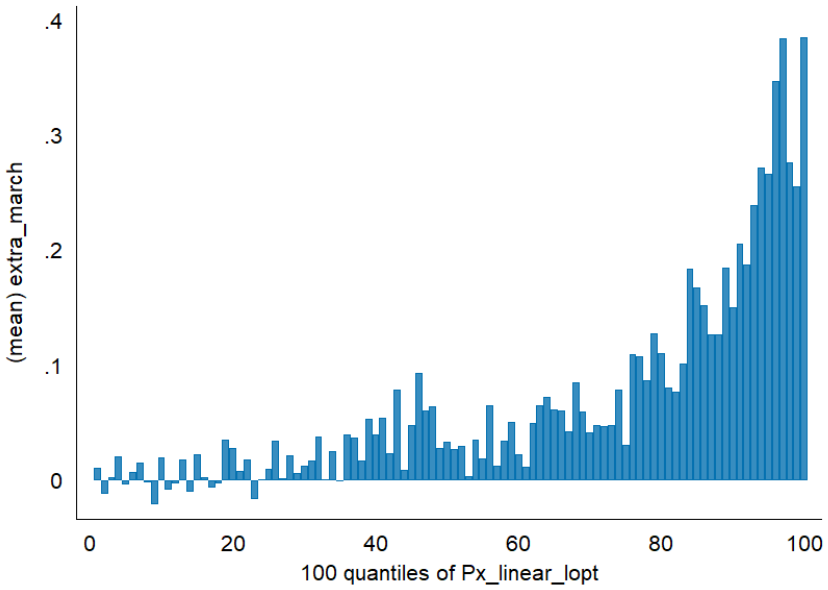
Table 5

Selected	*Lasso*	*Post-est lasso*
Sqmeter_inhabitant	0.00054	0.00126
elderly_alone	0.00042	0.00226
youth_unempl	-0.00013	-0.0003836
industrial_empl	0.00122	0.00140
artisan_empl	0.00013	0.00121
Public_mobility	-0.00035	-0.0027276
Mobility_slow	0.00083	0.00193
share_95_99	3.75164	6.39906
Mean_BAP	-0.01101	-0.0339027
4_quart_income	0.00812	0.02539
Partialled-out		
Spatial_lag_15min	0.00010	0.00012
Spatial_lag_30min	0.00002	0.00002
_cons	-0.06219	-0.1604717

Keep in mind that coefficients of Lasso should not be interpreted (Mullainathan and Spiess, 2017), but it is reassuring that the selected coefficients include the share of people between

95 and 99 years, since old people are known to be more at risk. An assessment of the training process can be to plot percentiles of the predicted values of the model with spatial controls and see what is the extra mortality rate. As expected, mortality rate starts from value around zero for low values of our predictor, and then increases sharply, especially in the top quartile.

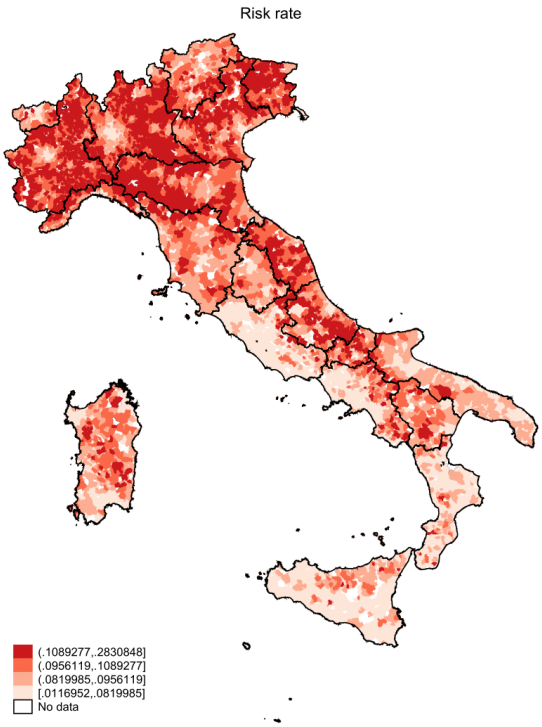
Figure 7



Analogously, one can estimate the extra death rate in the top quartile of our predictor and in the lower quartiles: such figures are 11.4% and 3.1% respectively. This means that our predictor does a good job in predicting which municipalities have a higher increase in mortality.

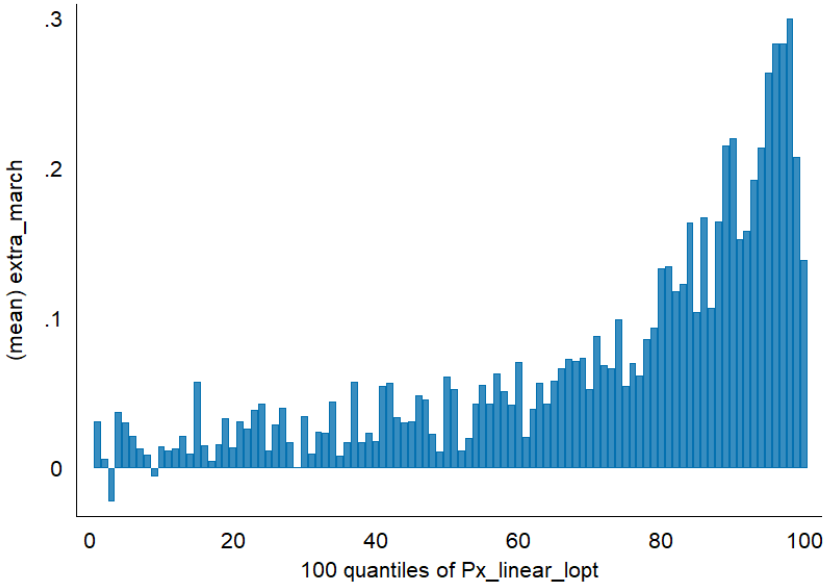
Then, we drop spatial predictors, and estimate the value predicted by the leftover coefficients. We call such value a "Covid risk rate". We can plot such risk over a map in Figure 8. It turns out that internal areas as well as country areas in the north are more at risk.
We can replicate the exercise with the second methodology, using the full training sample. Reassuringly, also when running on the full sample the lasso selects one of the spatial controls (even if it was not partialled out) and several demographic variables, although we obtain a

Figure 8



model with slightly lower predicting power. The predicted risk-rate is also very similar.

Figure 9

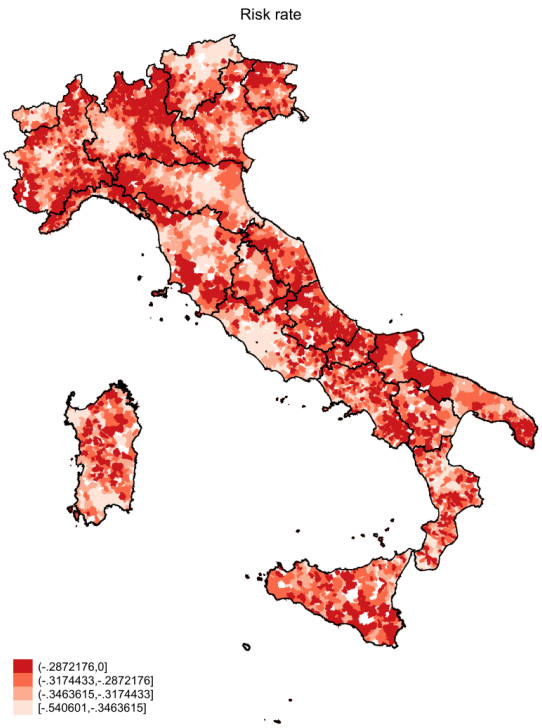


As a final check that our risk-rate is not trivial, we compare it with simpler predictors that

Table 6

Selected	*Lasso*	*Post-est lasso*
House_crowd_index	-0.0149	-0.0714
young_monoparental_household	-0.0071	-0.0270
lmkt_partec_women	-0.0010	0.0012
unemployment	-0.0012	-0.0054
low_skill_empl	0.0006	0.0052
Mobility_private	-0.0015	-0.0027
Share_60_64	0.0618	0.9390
Share_65_69	0.1735	1.5413
Share_85_89	1.1815	1.5697
Share_95_99	13.9974	18.3824
Mean_PM10	-0.0012	-0.0028
Beds_acute	-0.0004	-0.0101
5.q$_a vg_i ncome$	-0.0366	-0.0648
4.q_Daily_mobility 1	0.0300	0.0565
5.q_Outer_mobility	-0.0282	-0.0386
5.q_Public_mobility	-0.0035	-0.0617
Spatial_lag_30min	0.0002	0.0003
Partialled-out		
_cons	0.2475	0.0843

Figure 10



are used i the debate to define at-risk individuals. For example, if we use our methodology only in high-infection areas the R-square of our model is 15.4%, while if we use a model with only the share of people older than 65 plus spatial controls R-square is 8.5%.

7 Conclusion

In this paper we explored heterogeneities in the way Covid-19 impacted Italian municipalities. To the best of our knowledge, this is the first attempt to do so using a wide number of correlates, and including a machine-learning exercise. We exploited a newly assembled dataset with municipal level data provided by the Local Opportunities Lab, and we combined it with a dataset provided by ISTAT with the daily count of deaths in each Italian municipality. We used the excess morality (the gap between present and past deaths) as a proxy for Covid-related deaths. In order to see which features of a municipality correlate with the spread of the virus, we first implemented a triple-differences approach and then a cross-sectional first-differences approach. The results suggest that industrial employment share and average household dimension, are positively associated with Covid-related deaths and while average income, education, service employment share, trade employment share, public transport index, house crowd index, average house price and maximum level of NO2 are negatively associated. We want to stress that these are simple correlations, they must not be interpreted in a causal way and have to be read as suggestive evidence (as all the previous literature on this topic). We believe, however, that this is not a useless exercise. Firstly, correlations might narrow the range of possible causal explanations that further research will test. Secondly, the correlations we found are nonetheless useful to characterize the most at-risk areas of the country. With this spirit, we developed a machine learning methodology of prediction of higher increase in the death rates. We confirmed the insights that inner and peripheric areas are more at risk.

We conclude summarizing a few methodological lessons confirmed by this paper. First, granular variation in death rates and socio-economic variables is relevant, and comparison within highly infected regions is an important statistical tool, possibly better suited than

inter-regional comparison for getting credible insights. Second, when studying correlates of Covid19 with a broad prospective it becomes evident how difficult it is to avoid endogeneity and multicollinearity. Not surprisingly, Donald Rubin's slogan "No causation without manipulation" should be kept close in mind by researchers, especially when studying such a recent phenomenon as Covid19 epidemic. Third, a more appropriate and perhaps equally useful exercise - doable with municipality-level data - is a predictive one. In this article we proposed a methodology to use machine learning to predict areas at high risk of contagion independently from proximity to other beaten municipalities. The methodology is surely improvable, but the spirit is general and flexible. Our results are preliminary, but our hope is to point out a policy-relevant direction for research on the relation between socio-economic factors and Covid19 epidemic.

References

1. Alacevich, C., N. Cavalli, O. Giuntella, R. Lagravinese, F. Moscone, C. Nicodemo (2020). *Exploring the relationship between care homes and excess deaths in the Covid-19 pandemic: evidence from Italy.* IZA discussion paper, n. 13492

2. Aparicio, A. and S. Grossbard (2020). *Intergenerational residence patterns and COVID-19 fatalities in the EU and the US.* IZA discussion paper, n. 13452

3. Bartscher, A. K., S. Seitz, S. Siegloch, M. Slotwinski and N. Wehrofer (2020). *Social capital and the spread of Covid-19: Insights from European countries.* Covid Economics, Issue n. 26

4. Belloni, A., Chen, D., Chernozhukov, V., Hansen, C. (2012). *Sparse models and methods for optimal instruments with an application to eminent domain.* Econometrica, 80(6), 2369-2429.

5. Borgonovi, F. and E. Andrieu (2020). *Bowling together by bowling alone: Social capital and Covid-19.* Covid Economics, Issue n. 17

6. Borgonovi, F., E. Andrieau and S. V. Subramanian (2020). *Community-level social capital and COVID-19 infections and fatality in the United States.* Covid economics, Issue n. 32

7. Borjas, G. J. (2020). *Demographic determinants of testing incidence and Covid-19 infections in New York City neighbourhoods.* Covid Economics, Issue n. 3

8. Brandily, P., C. Brebion, S. Briole and L. Khoury (2020). *A Poorly Understood Disease? The Unequal Distribution of Excess Mortality Due to COVID-19 Across French Municipalities.* medRxiv.

9. Carillo, M. F. and T. Jappelli (2020). *Pandemic and local economic growth: Evidence from the Great Influenza in Italy.* Covid Economics, Issue n. 10

10. Carozzi, F., S. Provenzano and S. Roth (2020). *Urban density and Covid19.* IZA discussion paper, n. 13440

11. Chetty, R., Hendren, N., Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. American Economic Review, 106(4), 855-902.

12. Cole, M. A., C. Ozgen and E. Strobl (2020). *Air pollution exposure and Covid19*. IZA discussion paper, n. 13367

13. Desmet, K. and R. Wacziarg (2020). *Understanding spatial variation in covid-19 across the United States*. NBER working paper, n. 27329

14. Di Porto, E., P. Naticchioni and V. Scrutinio (2020). *Partial lockdown and the spread of Covid-19: lessons from the Italian case*.

15. Ding, W., R. Levine, C. Lin and W. Xie (2020). *Social distancing and social capital: why US counties respond differently to Covid-19*. NBER working paper, n. 27393

16. Dodanoglu, T. and E. Ozdenoren (2020). *Should I stay or should I go (out): The role of trust and norms in disease prevention during pandemics*. Covid Economics, Issue n. 16

17. Favero, C. (2020). *Why is Covid-19 mortality in Lombardy so high? Evidence from the simulation of a SEIHCR model*. Covid Economics, Issue n. 4

18. Galletta, S. and T. Giommoni (2020). *The effect of the 1918 influenza pandemic on income inequality: Evidence from Italy*. Covid Economics, Issue n. 33

19. Gerritse, M. (2020). *Cities and COVID-19 infections: Population density, transmission speeds and sheltering responses*. Covid Economics, Issue n. 37

20. Isphording, I. E. and N. Pestel (2020) *Pandemic meets pollution: poor air quality increases deaths by COVID-19*. IZA discussion paper, n. 13418

21. Kapoor, R., H. A. Rho, K. Sangha, B. Sharma, A. Shenoy and G. Xu (2020). *God is in the rain: The impact of rainfall-induced early social distancing on Covid-19 outbreaks*. Covid Economics, Issue n. 24

22. Knittel, C. R. and B. Ozaltun (2020). *What does and does not correlate with covid-19 death rates*. NBER working paper, n. 27391

23. Kuchler, T., D. Russel and J. Stroebel (2020). *The geographic spread of COVID-19 correlates with structure of social networks as measured by Facebook*. NBER working paper, n. 26990

24. Moller, C., C. W. Hansen and P. T. Jensen (2020). *The 1918 epidemic and a V-shaped recession: Evidence from municipal income data*. Covid Economics, Issue n. 6

25. Sà, F. (2020). *Socioeconomic determinants of Covid-19 infections and mortality: Evidence from England and Wales*. Covid Economics, Issue n. 22

26. Schmitt-Grohé, S., K. Teoh and M. Uribe (2020). *Covid-19: Testing inequality in New York City*. Covid Economics, Issue n. 8

27. Sussman, N. (2020). *Time for bed(s): Hospital capacity and mortality from COVID-19*. Covid Economics, Issue n. 20

28. Tortuga (2020). *Fase 2: Sistemi locali del lavoro*. `https://www.tortuga-econ.it/wp-content/uploads/2020/05/SLL-REPORT-WORD-aggiornato-10-aprile_v2.pdf`

29. Tubadji, A., D. J. Webber and F. Boy (2020). *Cultural and economic discrimination by the Great Leveller: The COVID-19 pandemic in the UK*. Covid Economics, Issue n. 13