Human mobility and COVID-19 epidemic

Andrea Nicolai, Karan Kabbur Hanumanthappa Manjunatha, and Camilla Quaglia

Abstract– With the onset of COVID-19, a lot of research focused mainly on modelling the epidemic. However, earlier research indicates that mobility is also a crucial factor in spread of epidemic [11]. So, for COVID-19, in our work we investigate how Italian mobility data obtained from Google repository could be related to trend in epidemic spreading namely the prevalence using statistical tools such as Kendall τ. Another part of our work includes estimating R_{eff} using MCMC technique and observing whether there is a relation with respect to transit mobility. In the end, we see that mobility can be a proxy to human encounters and generally describe epidemic spreading but with certain exceptions, finally we find that an increase in transit mobility can lead to increase in R_{eff} in non-linear fashion.

INTRODUCTION

In the last year, COVID-19 has definitively changed everyone's life. Therefore it suddenly became one of the most important topics scientific research focused on, in all its facets. For instance, the discussion mainly dealt with clinical and medical data, but later the epidemiological modeling of diseases started to play a role: it helped administrations to allocate enough resources to efficiently face the predicted number of cases and hospitalized people, as well as the risk assessment useful to enact the correct travel restrictions and policies. Due to much more and increasing amount of data collected every day from mobile phones and GPS, mobility data has increased its relevance in the last pandemic and has been started to be studied deeper [8]. Indeed, in response to the outbreak, Google published mobility data [4] collected from users of its mobile applications.

In this work we examine Italian data, in particular at a regional resolution. It allows to further investigate the eventual relationships between mobility and COVID-19 spreading, in particular to investigate the effects of government restrictions and health advice on the outbreak.

In first part, we try to spot any correlation present between prevalence and mobility curves for every Italian region [21], thus estimating the lag in days where the value for Kendall's τ correlation coefficient is maximal. In second part, we try to estimate the daily effective R_{eff} for every region using a MCMC [22], finally investigating whether R_{eff} and mobility are somehow related by the means of a non parametric fit.

METHODS

Data collection

Data concerning the prevalence of the disease for each Italian region are retrieved from the following repository [13]. The prevalence in particular is obtained subtracting to the *'Confirmed'* column in the dataset (that represents the cumulative number of confirmed cases), *'Deaths'* (cumulative number of deaths) and *'Recovered'* (cumulative number of recovered). The incidence, i.e. daily new positives, is instead extracted from the following dataset [1]. As anticipated in the introduction, Google Community Mobility Reports allow to study the relationship between mobility and COVID-19 outbreak. The data shows the change in terms of visitors to the categorized locations or time spent at those locations compared to reference days. A reference day represents a normal value for that day of the week. The reference day is the average value for the five-week period from January 3rd to February 6th, 2020. Data are divided into different categories, depending on the place the destination of the journey: *retails and recreations*, *parks*, *stations*, *workplaces*, *residential areas* and *groceries and pharmacies*. Regarding mobility dataset it should be further noted that grocery and pharmacy are both combined as these tend to be considered essential trips, while parks and transit stations refers to a number of places put together:

Parks	Transit stations
Public garden	Subway station
Castle	Sea port
National forest	Taxi stand
Camp ground	Highway rest stop
Observation deck	Car rental agency
Public beaches	Railway and bus stations

To perform this analysis both epidemiological and mobility datasets were filtered, removing the information on days when the mobility reports do not track any data.

In fact if the activity was too low in a specific day, missing values were reported, indicating that it was below the threshold of anonymity set by Google.
One last remark is that one has to take into account that the Google's data are based on information from users who have turned on the 'location history' option of their account. Therefore this information may not represent the precise behavior of the whole population.

Kendall's τ

To study the correlation between mobility and epidemiological data the Kendall's τ is calculated, as a correlation measure. It varies in a range [-1,1], where values close to 1 indicate strong agreement while values close to -1 indicate strong disagreement. The reason behind the choice of this kind of rank correlation coefficient is the non-parametric nature of data, given that the Kendall's τ does not depend upon the assumptions of various distributions. Moreover using this coefficient, *p-values* are more accurate with smaller sample sizes, rather than others non-parametric rank coefficients, such as Spearman's ρ [9].

In the computation the variant considered is the τ_b defined as

$$\tau_b = \frac{P - Q}{\sqrt{(P+Q+T) \cdot (P+Q+U)}} \tag{1}$$

where P is the number of concordant pairs, Q the number of discordant pairs. Called x and y the vectors of rankings, T represents the number of ties only in x, and U the number of ties only in y. If a tie occurs for the same pair in both x and y, it is not added to either T or U [15]. Kendall's τ rank correlation is computed for mobility and prevalence by implementing `kendalltau()` function from `scipy.stats` module in *Python 3*.

Our expectations: The experimental hypothesis is that reductions in travels, as well as mobility increasing towards residential places, will lead to fewer contacts among people thus lowering the rate of transmission of the virus and hence should bring also reductions in the incidence of the epidemic. Moreover, from our analysis, we expect a positive correlation between prevalence data and mobility toward residential places, as well as negative correlation when considering the transit mobility. These two aforementioned mobility categories are taken into account to produce the results, out of all the categories of Google's mobility reports data. This choice is motivated by the fact that mobility related to transit and residential places is the most relevant from the epidemiological point of view, as a preliminary analysis stated.

The calculation of the Kendall's τ is performed using Google's mobility and prevalence data. The latter is considered after having shifted it backwards of 16 days, and only then was related to mobility. As an example, we want to spot correlations between mobility data of day April 1st with prevalence data available on April 17th, mobility data on April 2nd with prevalence on April 18th etc. To justify the presence of such lag, chosen as the best one in the result section, one has to consider that the incubation period for COVID-19, which is the time between exposure to the virus and symptom onset, can be up to 14 days (on average 5-6 days). Therefore, even before analysing data, one would expect a shift of few more days with respect to 14, due to delay in testing and reporting.

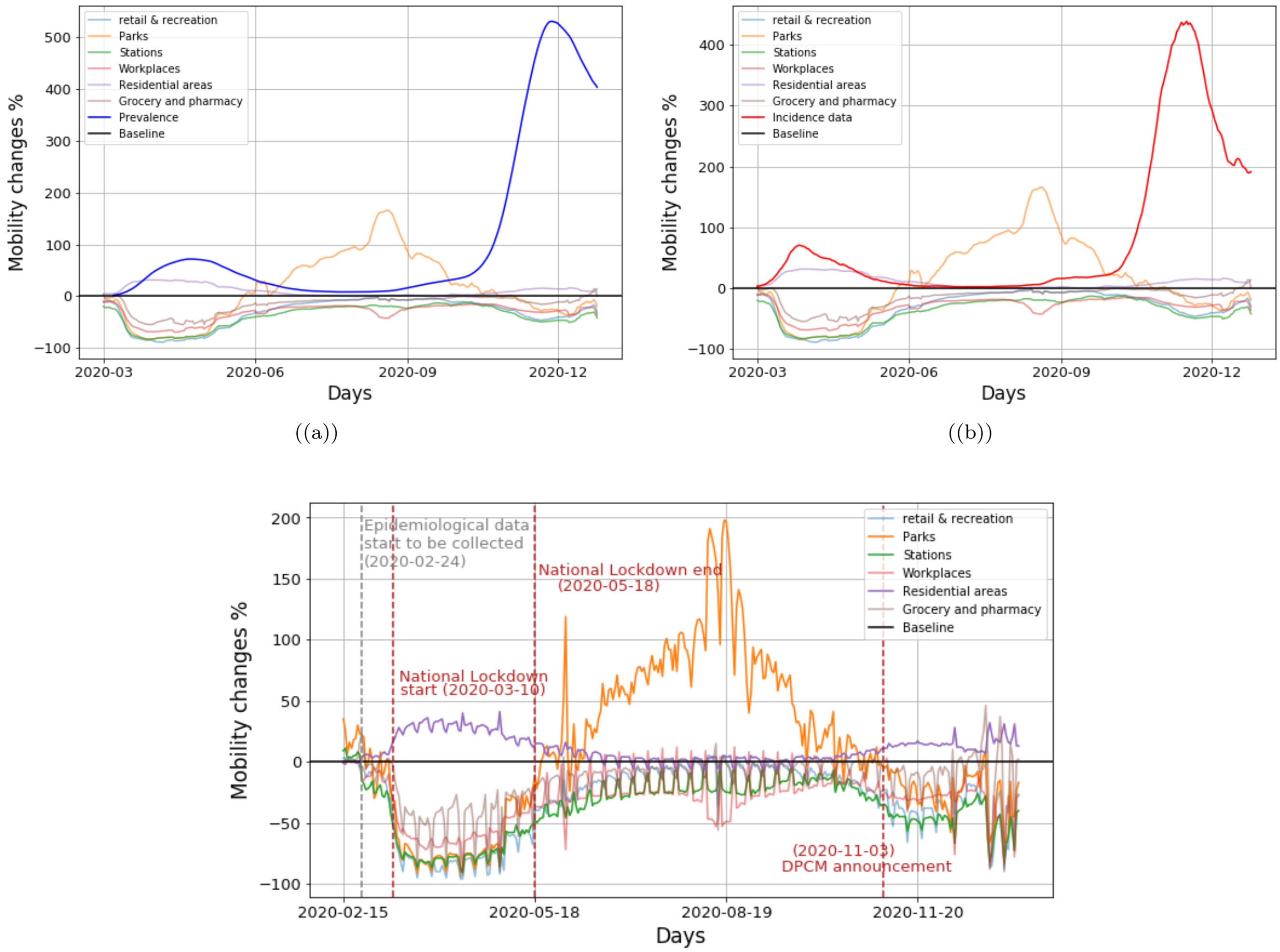
The analysis is performed in three different periods, marked in Fig. 1(c), the '1st lockdown' (from 2020-03-10 to 2020-05-18), the 'summer period' (from 2020-05-18 to 2020-11-03) and the '2nd lockdown' (from 2020-11-03 to 2020-12-25). It is important to state that we refer to the period where there were no mobility restrictions as "Summer". Indeed, the choice of this time division is clearly driven by different mobility restrictions the government emanated. In the latter period the analysis is slightly different. In fact on 2020-11-03 the Italian government emanates a DPCM that divides the national territory in different risk areas, characterized by different mobility restrictions. The ordinance provides for three different colors, red, orange and yellow for the Italian regions, corresponding to very strict, severe and blind mobility restrictions. Such classification also means zones with 'maximum' risk of contagion for 'Red', 'elevated' risk for 'Orange' and 'moderate' risk for 'Yellow' whose restriction rules based on color codes for various regions can be found in [6]. Data for the different "colors" of region have been manually scraped from [19].

R_{eff} estimation

From theory we know that at every timestep, which in our case is a single day long, the actual number of new infections y_t is on average function of the number of infected individuals at previous time steps $y_{t-1}, y_{t-2}, ...$, times the so called *effective reproduction rate* $R_{eff}(t)$, according to the following convolution:

$$y_t = \sum_{i=1}^{M} y_{t-i} \cdot R_{eff}(t-i) g_i \tag{2}$$

where we also introduced g_i, that is the generation time distribution for the delay i. This formalism comes from [18], where we take into account that the probability of a person to infect someone else depends on time (infec-



((a))

((b))

((c))

FIG. 1:
(a) Behaviour of the National prevalence curve compared with Google's mobility data. The period displayed is from 2020-03-01 to 2020-12-25. To both data a rolling average with 7 days window was applied. Prevalence is rescaled so as to plot in the same axes as mobility curves which helps in better comparison.
(b) Behaviour of the National incidence curve compared with Google's mobility data. The time interval is the same as in graph (a). Incidence is rescaled such that to make the curves comparable. As in graph (a) the rolling average with the same window is applied.
(c) Google's mobility raw data. They show a strong periodicity, especially in some categories, due to weekends or festivities. Bounds for the three phases we have defined, namely first (from 2020-03-10 through 2020-05-18) and second lockdown (from 2020-11-03 through 2020-12-25) and summer (from 2020-05-18 through 2020-11-03), are visible.

tivity indeed is time dependent), and we assume that serial intervals distribute similarly to generation times. In other words, we treat serial interval as a proxy for the generation time.

Up to now we have discussed the way infected individuals transmit the disease to other people and the latter ones becomes in turn infected, however, our observable is only the number of positive tests: we do not have any data about when the disease was transmitted. We refer to the time elapsed between the infection and the positive test as delay distribution [23]. In order to estimate this distribution, the difference in days between the date of onset of symptoms and the date of the positive test was selected, plus the incubation time that is assumed to be exactly 5 days [17].

The convolution between the number of people that contract the infection for every day, and the just mentioned delay distribution, returns us the distribution of the number of positive tests for a certain day, given today's situation.

There is another adjustment before proceeding to the estimation of $R_{eff}(t)$. It is pretty straightforward that the number of positive tests depends on the number of tests that are performed: the more tests are made, the larger the number of people tested positive. This variability has to be taken into account, since there might be strong fluctuations in the number of tests processed due to weekends or holidays resulting in a biased R_{eff}. Hence in our model we multiply e_t, a normalized quantity proportional to the number of tests performed, by the number of positive tests described by the generative process z_t. The expected number of positive tests $\tilde{z}_t$ is:

$$\tilde{z}_t = e_t \cdot z_t \tag{3}$$

In other words thanks to testing exposure e_t, if the tests made were more the expected number of test positive results would be more. It is a very similar concept to the exposure parameter in Poisson regression.

The algorithm to estimate the number of secondary cases R_t originating from infected people at time t is the following:

- Primary infection occurs at time t, this is our starting point and R_t refers to this moment
- Generation time elapses until a secondary infection occurs, according to the first distribution we introduced that includes also the infectivity depending on time
- Onset time passes until secondary infections have occurred and the individuals develop symptoms, finally being tested positive. This would be the number of positive tests if number of tests processed were assumed to be constant every day.
- Multiply the latter number of positive tests by the testing exposure e_t

The output of this procedure is the quantity we will use to fit the data and estimate $R_{eff}(t)$, by the means of a MCMC that sequentially iterate over the dates. Prior of $R_{eff}(t+1)$ for next date $t+1$ is assumed to be a random walk starting from $R_{eff}(t)$, so to avoid big jumps.

For the simulation [22] was used that automatizes the whole just mentioned algorithm, that is in turn based on [20].

Non parametric fits

To answer the question ”Are variation in mobility related to variations in epidemic spreading potential or speed (exponential growth, R_{eff} or R_t)?”, a scatter plot of R_{eff} vs transit mobility(% change from baseline) is done where R_{eff} was obtained through the procedure as explained in previous sub-section. With this scatterplot, it enables visual assessments of relationships or functional dependencies between the two variables namely R_{eff} and mobility included in the display. However, the problem with this is noisy data values, sparse data points in some parts in the scatterplot and weak interrelationships could inhibit visual identification of any such patterns. Thus, Loess fitting comes to the rescue. *Jacoby* in his paper on *Electoral Studies(2000)*[14] gives a pragmatic understanding on how the fitting works and what the fitting means.

Loess is an acronym for *locally weighted regression.* It is nonparametric fitting technique that does not require a priori specification of the relationship between the dependent and independent variables. In contrast with parametric fitting which instead summarizes the relationship between the dependant and independent variables, when the structure in the data conforms to the type of function that is fitted by smoothing algorithm. However, when this ”correct” functional form is almost always unknown, loess non parameteric fit is a better option.

Loess has always been used as a descriptive and exploratory tool for fitting smooth curves to scatter plots and is employed most frequently for these kinds of situation. The non parametric smoothers such as this, is used to locate smooth curve among data points without requiring any advance specification of functional relationship between the variables unlike parametric fitting. The fitting algorithm simply tries to follow the empirical concentration of the plotted points. The resultant fitted “line” should pass through the most dense areas of the data region in the scatterplot, regardless of the shape of the curve that is required in order to do so.

Since with the visual inspection of scatter plot alone is problematic in discerning the precise nature of dependence of the variables, fitting a smooth curve to the points in scatter plot is a good strategy to deal with that. The purpose of the curve is to summarize the central tendency of the Y variable’s distribution at different locations within the X variable’s distribution. If the two variables are unrelated to each other, then the smooth curve will be a flat line (the center of the Y distribution does not change, regardless of the X value). If the two variables are related, then the smooth curve should exhibit some other, non-horizontal shape.

RESULTS

Kendall’s τ results

Firstly, the results that state which is the best lag in days between epidemiological data and mobility are shown. In order to compute the lag, which will be the same regardless of the period (periods are shown in Fig. 1(c)) and of the region, we used National level data of the first lockdown. The plots in Fig. 5 show how Kendall’s

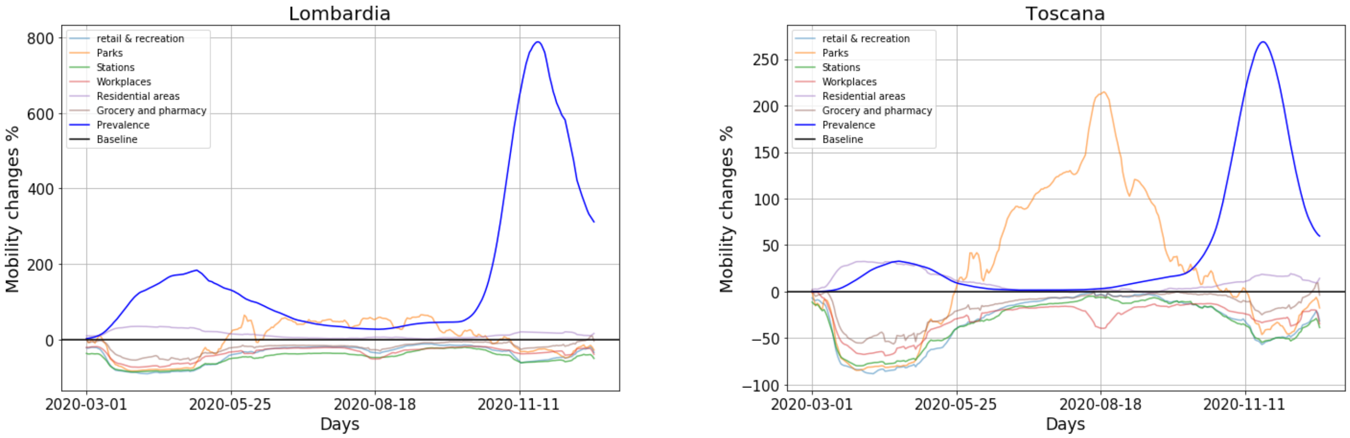


FIG. 2: Behaviour of the prevalence curve compared with Google's mobility data for Lombardy and Tuscany. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility.

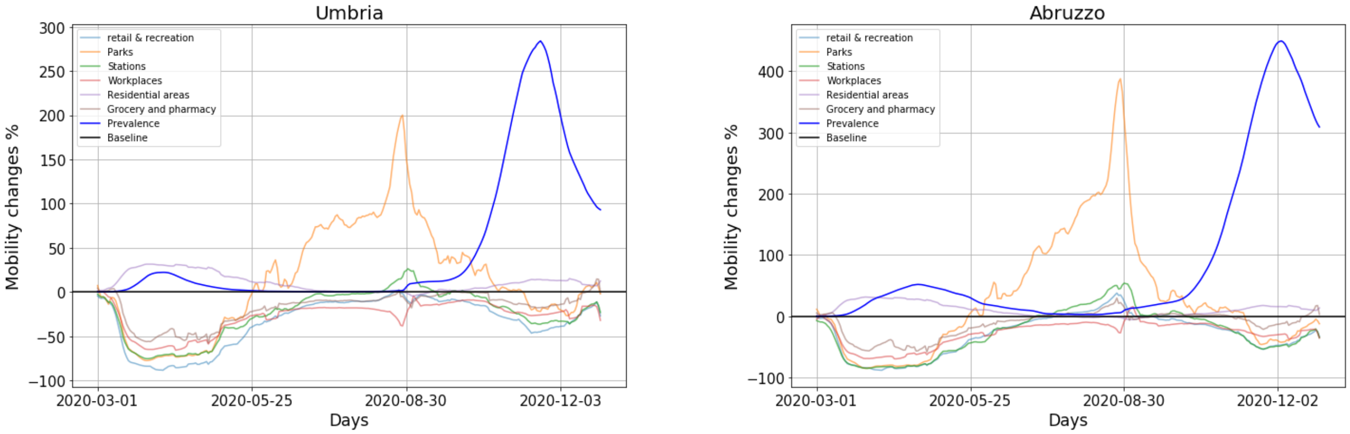


FIG. 3: Behaviour of the prevalence curve compared with Google's mobility data for Umbria and Abruzzo. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility.

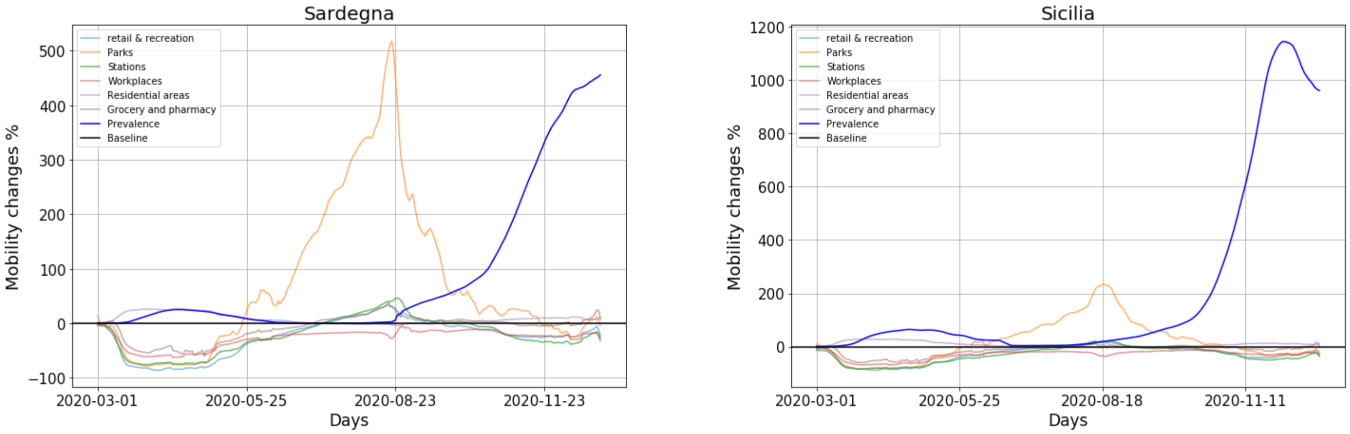


FIG. 4: Behaviour of the prevalence curve compared with Google's mobility data for Sardinia and Sicily. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility.

τ varies in function of the lag in days, considering both incidence and prevalence data, for all the mobility categories. All the variables, namely incidence, prevalence and mobility were filtered using a rolling average of 7 days window due to the periodicity appearing at weekends as well as to minimize spurious fluctuations in the data collected. We recall that the aim was to find the lag such that Kendall's τ was the largest. This is obtained for a lag around 16 days, as one can clearly see in Fig. 5, left plot, using prevalence data. Therefore, from now on, prevalence will be used when computing the τ correlation, having applied the best shift we have just found.

However, there is an interesting pattern to be noted in the right plot: correlation with incidence is indeed maximum when no shift is applied and it decreases as the lag becomes larger. After 16 days, the curve flattens.

The analysis in the 1st lockdown (from 2020-03-10 to 2020-05-18) and in the summer period (from 2020-05-18 to 2020-03-11) is carried on with a similar approach. In each of the two aforementioned phases a time resolution of 20 days is chosen, with the prevalence data shifted with respect to the mobility ones by 16 days. The last time slot in the summer period is chosen of 30 days instead of 20, in order to arrive at the date of November DPCM, after which the analysis changes because of the different coloured regions. In the investigation, 'Trentino Alto Adige' region is neglected, since the epidemiological data were given separately for 'Trento' and 'Bolzano', while the mobility is for the whole region. Since in the 2nd lockdown (from 2020-03-11 to 2020-12-25) the two provinces are coloured differently (different mobility restrictions), it has been decided to neglect this region from the whole analysis because of too many complications.
Moreover in the summer period the 20-days time slot from 08-06 to 08-26 is missing, due to lack of points within the interval: for some days data was not available. Kendall's τ correlations between prevalence and mobility, respectively for the residential and transit, for 1st lockdown and summer are shown in Fig. 6 and 7. Note that the mentioned boxplots group all the Kendall's τ values that were computed for each region. Moreover, the result for single regions that are colored according to their Kendall's τ values is shown, for residential and transit mobility in the 1st lockdown, respectively in Fig. 8 and 10. Whereas the same maps, but referring to Summer period are, respectively for residential and transit mobility, the one shown in Fig. 9 and 11.

For second lockdown period, it was decided to divide the whole time interval into two phases, respectively phase 1 (November 03rd - December 04th) and phase 2 (December 05th - December 25th). Phases were chosen with such long time period in order to have enough data to run a meaningful statistics (i.e. Kendall τ computation) on it as well as a new DPCM came into effect on 4th of December[5] thus restrictions/colors of regions again changed, hence that breaking of 2nd lockdown on that particular day into 2 phases is pragmatic and very much inline with our research. Moreover, for each phase the amount of days a single region had either yellow, orange or red restrictions on mobility was noted by observing from [19] and finally assigned the color for the whole length of each phase seperately according to the largest amount of days a region had certain restriction/-color. Results of this labelling are shown in Table I.

Region	**1st phase**	**2nd phase**
Piemonte	*red*	*yellow*
Valle d'Aosta	*red*	*orange*
Lombardia	*red*	*yellow*
Veneto	*yellow*	*yellow*
Friuli Venezia Giulia	*orange*	*yellow*
Liguria	*orange*	*yellow*
Emilia-Romagna	*orange*	*yellow*
Toscana	*red*	*orange*
Umbria	*orange*	*yellow*
Marche	*orange*	*orange*
Lazio	*yellow*	*yellow*
Abruzzo	*red*	*orange*
Molise	*yellow*	*yellow*
Campania	*red*	*orange*
Puglia	*orange*	*yellow*
Basilicata	*orange*	*yellow*
Calabria	*red*	*yellow*
Sicilia	*orange*	*yellow*
Sardegna	*yellow*	*yellow*

TABLE I: Regions and their color in the second lockdown for different phases. First phase is from November 03rd through December 4th, while the second phase starts from December 5th through December 25th.

Carrying out the same analysis as before, that is to say computing Kendall's τ for every single region but now grouping them according to their color, the results one obtain are shown in Fig. 12. Groups with a number of samples less than 5 are shown as distinct points and without the boxplot, since Krzywinski *et al* in there $Nature$ report [16] mentions that box plots are to avoided for very small samples($n < 5$). Moreover, Italian choropleth maps with region colored according to Kendall's τ and different mobilities are shown in Fig. 12 and for visual comparison, beside them are also regions colored according to the restrictions that they are placed in those respective phases.

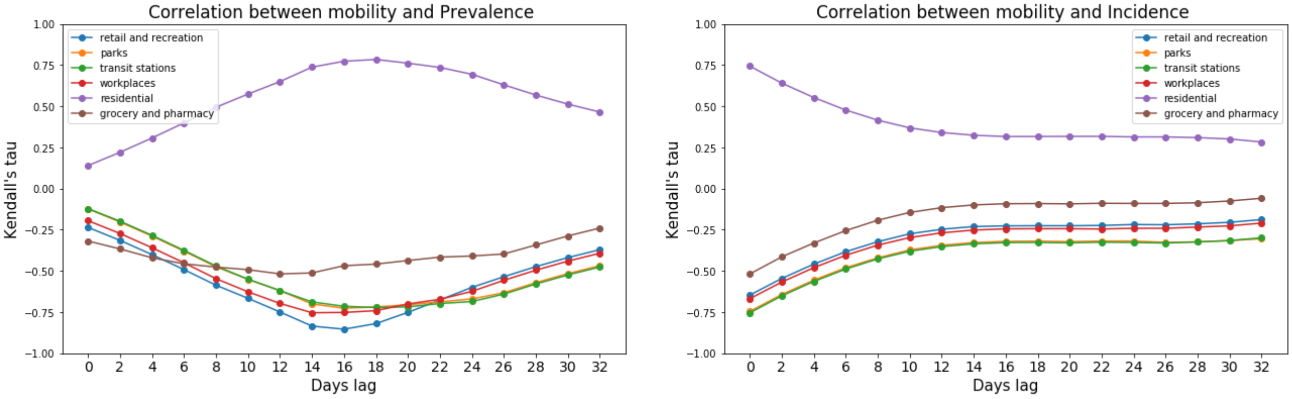


FIG. 5: Kendall's τ coefficient between prevalence (left) and incidence (right) and the mobility in all the categories provided by Google. The former is plotted vs the lag applied to epidemiological data to see empirically which is the best shift.

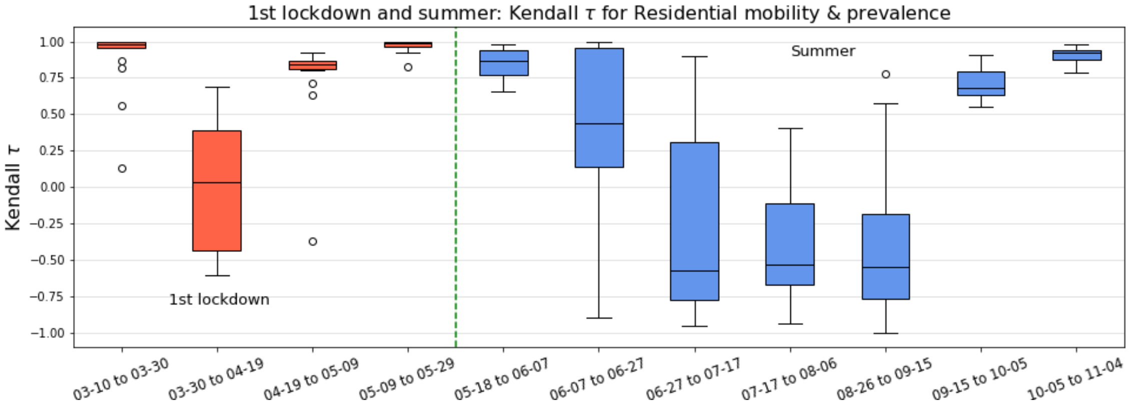


FIG. 6: Kendall's τ coefficient for 1st lockdown and summer period for residential mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days). Data for all regions were considered.

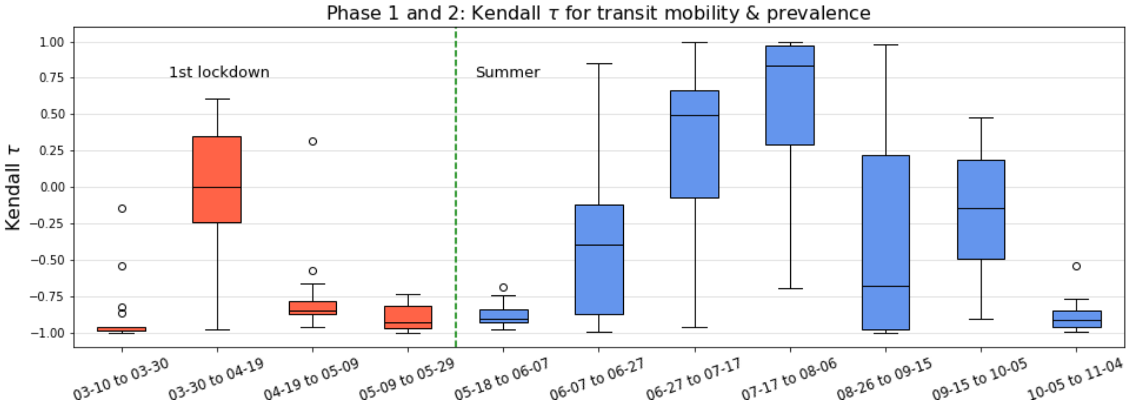


FIG. 7: Kendall's τ coefficient for 1st lockdown and summer for transit mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days). Data for all regions were considered.

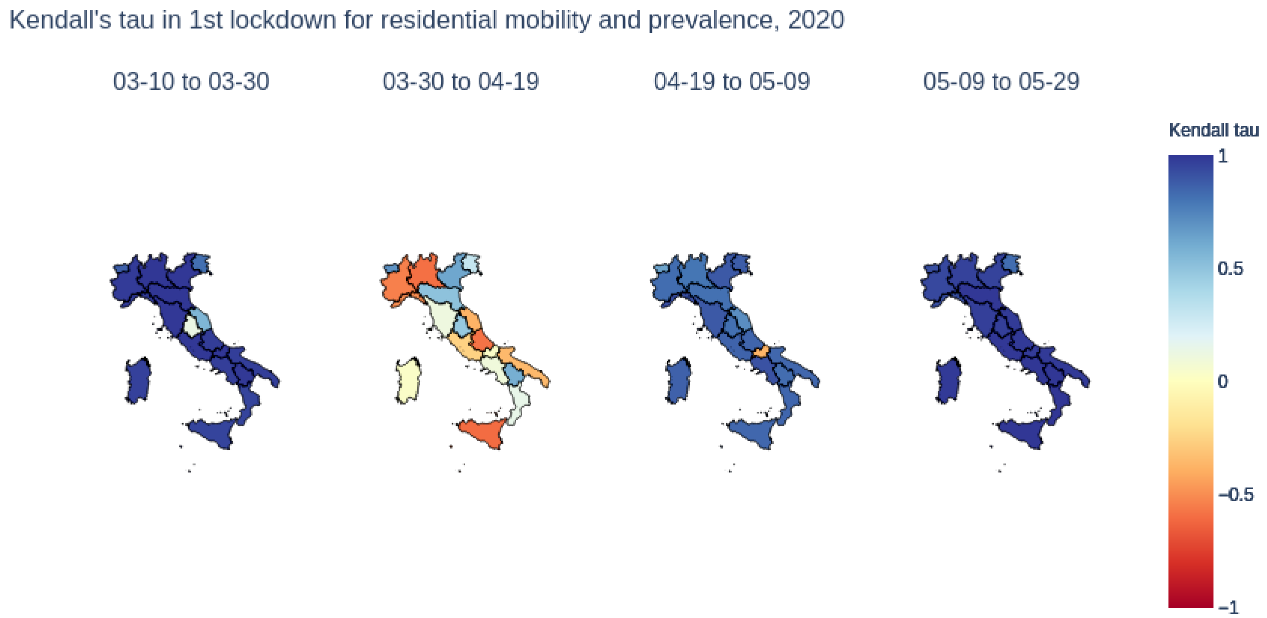


FIG. 8: Choropleth map showing Kendall's τ coefficient for 1st lockdown for residential mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days).

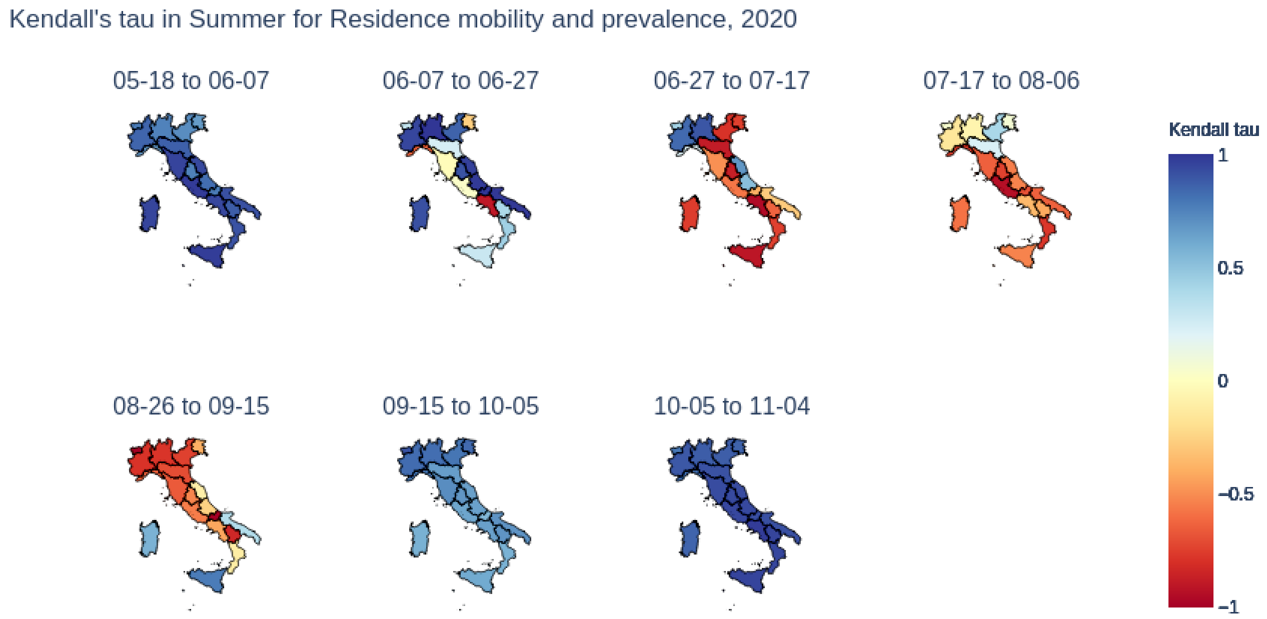


FIG. 9: Choropleth map showing Kendall's τ coefficient for Summer period for residential mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days).

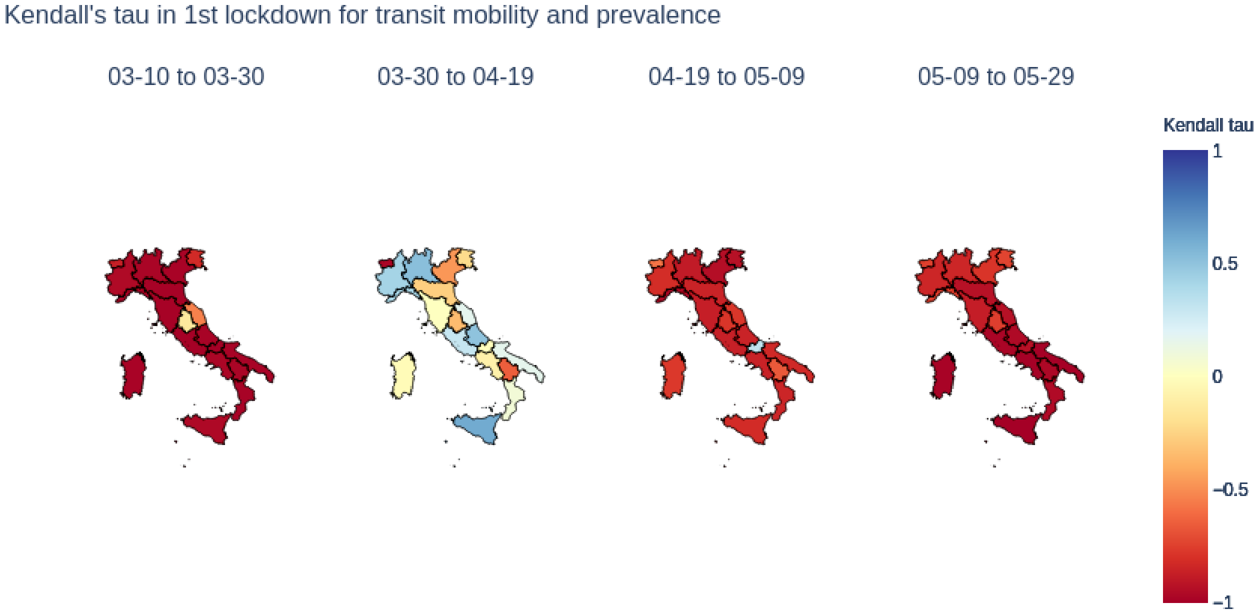


FIG. 10: Choropleth map showing Kendall's τ coefficient for 1st lockdown for transit mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days).

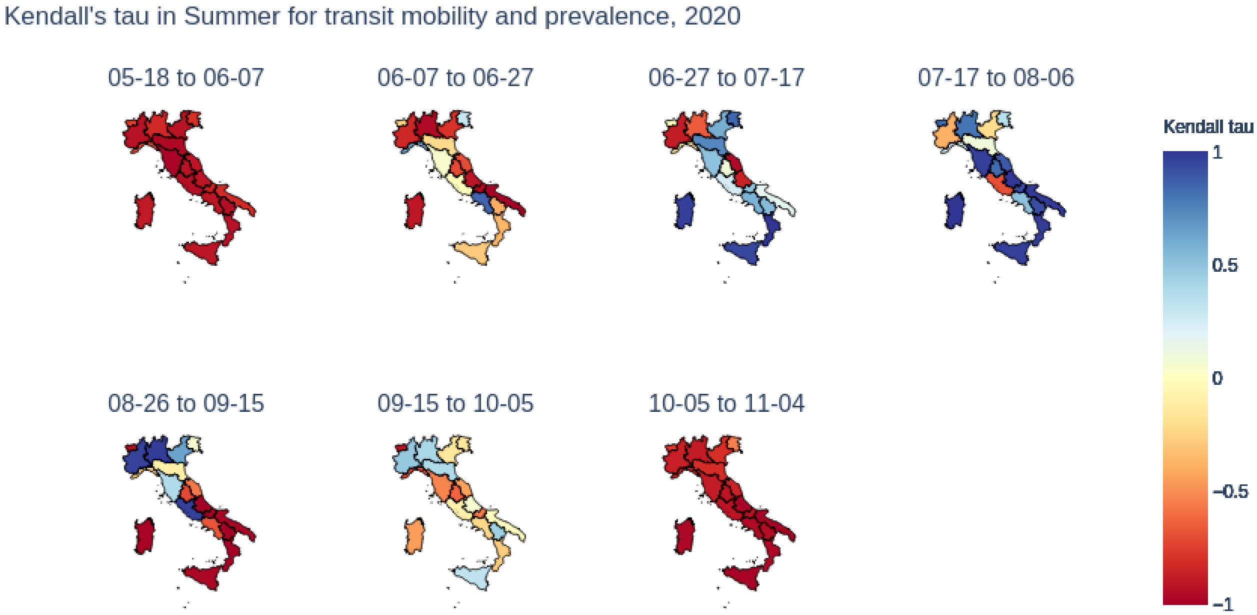


FIG. 11: Choropleth map showing Kendall's τ coefficient for Summer period for transit mobility provided by Google and prevalence shifted backwards by the optimal lag (16 days).

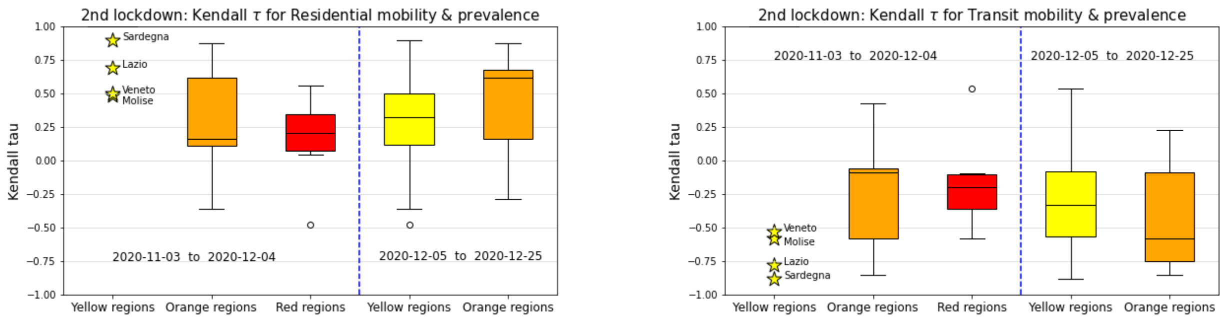
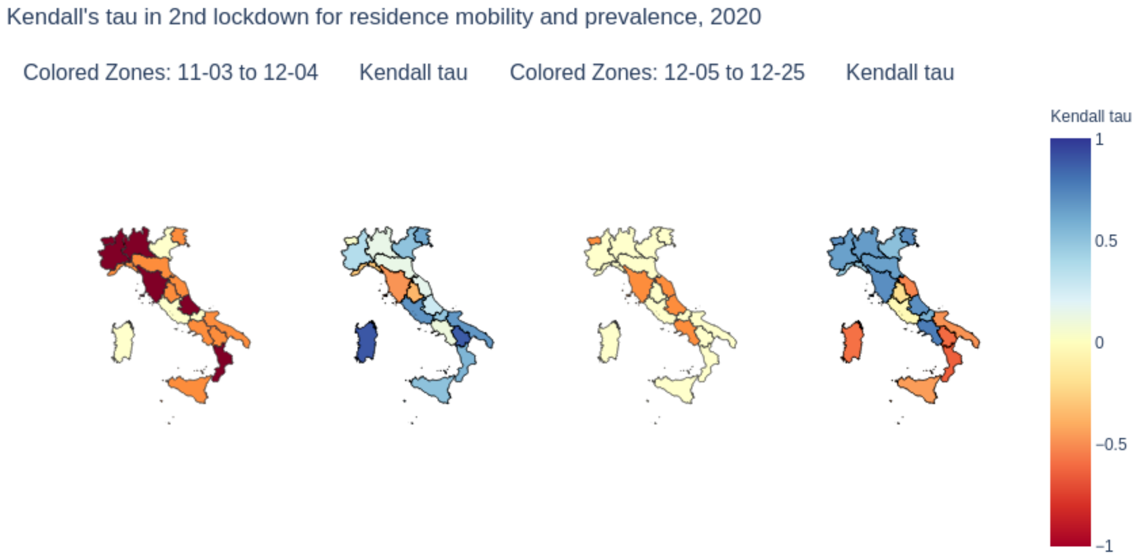
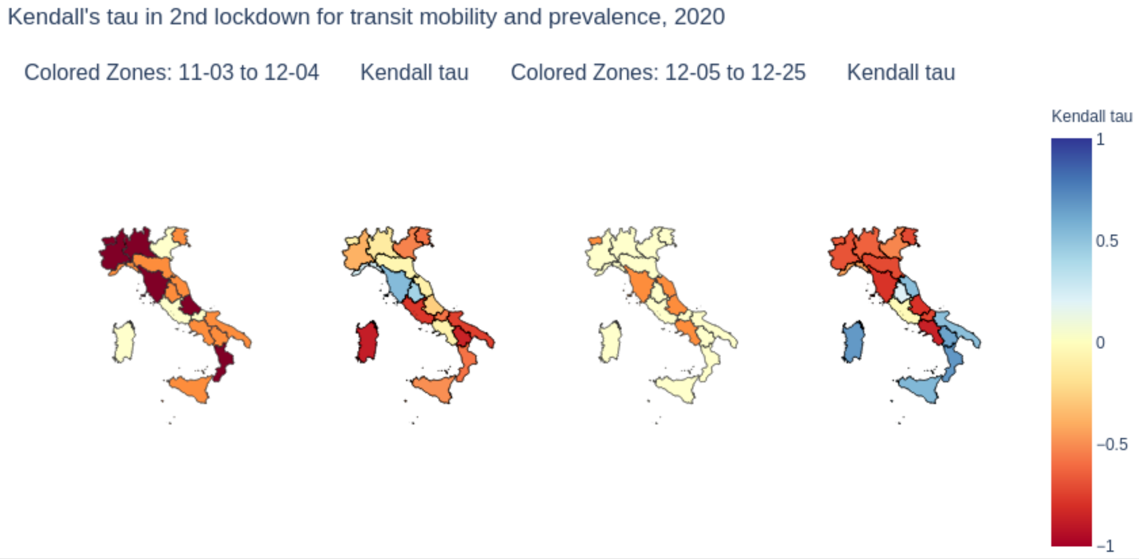


FIG. 12: **Left.** Kendall's τ coefficient for the 2nd lockdown period and residential mobility, with regions grouped according to the actual mobility restriction at that time.
Right. Kendall's τ coefficient for the 2nd lockdown period and transit mobility, with regions grouped according to the actual mobility restriction at that time.


Kendall's tau in 2nd lockdown for residence mobility and prevalence, 2020
Colored Zones: 11-03 to 12-04
Kendall tau
Colored Zones: 12-05 to 12-25
Kendall tau
Kendall tau
1
0.5
0
−0.5
−1


Kendall's tau in 2nd lockdown for transit mobility and prevalence, 2020
Colored Zones: 11-03 to 12-04
Kendall tau
Colored Zones: 12-05 to 12-25
Kendall tau
Kendall tau
1
0.5
0
−0.5
−1

MCMC simulations

For every region, four chains were run. The number of burn-in samples and the total number of draws were respectively 100 and 150 for each of the chains. For Italy, considered as a whole country the output was the one shown in Fig. 13. One can distinguish the three periods our analysis is based on, which are highlighted by vertical dashed lines: first lockdown, summer period and second lockdown.

In addition we present in Fig. 14 the R_{eff} for 4 other regions: Piemonte, Lombardia, Veneto and Friuli Venezia Giulia. One can easily note the bounds for the different time phases: 1st, 2nd lockdowns and summer period. It is shown and highlighted by stars also initial and final R_{eff}, this would help in consistency check for data in Fig. 16. The R_{eff} vs t plots for other regions is presented in the appendix Fig. A.5.

R_{eff} vs Transit mobility

Once obtained the $R_{eff}(t)$ for every region, we performed the Loess non parametric fit for the scatter plot of R_{eff} vs Transit mobility following two different approaches. `lowess()` function from `statsmodels.nonparametric.smoothers_lowess` module of `Python` was used to do this fit.

The first one by creating a scatter plot containing transit mobility data for all the regions on x-axis and $R_{eff}(t)$ for all the regions on y-axis. A single point represents, in other words, the pair:

$$(\text{mobility transit } (t), R_{eff}(t))_i$$
$$i = \text{Lombardia, Veneto, Campania...}$$

for a given day t. The resulting fit is the one shown in Fig. 15. Red square markers refers to 1st lockdown, blue triangle markers refers to Summer period/after the lifting of 1st lockdown and finally green circles refer to 2nd lockdown. Intensity of the color, which could be either red, green or blue, according to the phase they refer to, increases in accordance with the time passed from the start of each phase: the darker, the later it is with respect to time for that phase. In this way we can represent time dimension as well.

The second approach we followed was to produce a series of scatter plots one for every region. In such plots, as before, on x-axis we represented the transit mobility, while on y-axis the R_{eff}, and points become darker as time elapses from the starting of each phase. Only four regions are represented here in Fig. 16, namely Veneto, Lombardia, Piemonte and Friuli Venezia Giulia, remaining ones are in the appendix Fig. A.4.

DISCUSSION

Kendall's τ lag. According to Fig. 5, one can hypothesize that at least after 16 days the mobility restrictions have effect, and that is why curve flattens in the right plot for larger lags. This can be true also because considering average generation times, incubation period and eventual delay in reporting cases might last up to 16 days. This is in line with the reports where the incubation period for COVID-19 is thought to extend to 14 days, with a median time of 4-5 days from exposure to symptoms onset[12][7]. One study reported that 97.5% of people with COVID-19 who have symptoms will do so within 11.5 days of SARS-CoV-2 infection.[7]

1st lockdown. Regarding the first lockdown the general behaviour at the National level can be inferred by the plot in Fig. 1(a). A similar plot, referred to each Italian region, is done and turned out to be useful for this discussion. As the map in Fig. 8 confirms, in the very first time slot from 03-10 to 03-30 the correlation between prevalence and residential mobility is strongly positive, for the majority of regions, since we have an increasing behaviour in both curves. In the same period the mobility related to transit decreases, and this brings a strong negative correlation with respect to prevalence, as Fig. 10 states. However two outliers regions are spotted clearly from maps: Abruzzo and Umbria. For the latter we see almost zero correlation. Indeed their prevalence curve is still very low, as we can see from Fig. 3. As one can infer from the plots that at time the mobility both for transit and residential stay almost flat for a while. Despite this, in the period from 03-30 to 04-19 a different behaviour of the Kendall's τ among regions is shown in both Fig. 8 and 10. It reflects also in a wider boxplot in Fig. 6 and 7 (red parts). E.g. There are regions, like Toscana (Fig. 2), that show almost no correlation, due to the fact that the prevalence curve remains almost flat for a while, too. Also regions like Sicily and Lombardy show a positive correlation among prevalence and transit and a negative one with the residential mobility. However during this period we are on the peak of the prevalence curve. Indeed the boxplots in Fig. 6 and 7 (red parts) show a median that is close to zero in the corresponding time slot. For the two others periods (04-19 to 05-09 and 05-09 to 05-29) we see again a positive correlation among prevalence and residential and a negative one if transit is considered. This is due to the fact that some mobility restrictions are gradually released [2] [3] and the prevalence curve begins to decrease. Indeed people started to go out more. Therefore residential and prevalence behave in the 'same way', differently from transit mobility. So, the correlation (either positive or negative) is almost at the extremes (1 and -1) in the last period and is almost

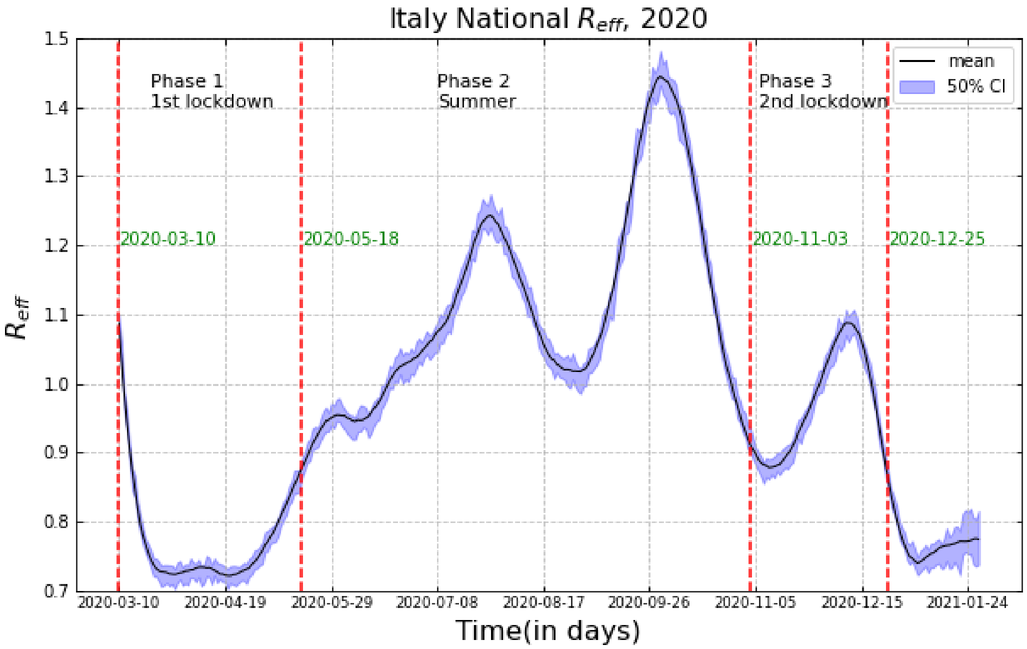


FIG. 13: Trend of Italian R_{eff} wrt time, estimated by the means of a Markov Chain Monte Carlo. Vertical dashed lines represent the different periods considered in our analysis, referring to first lockdown, summer time and second lockdown until Christmas day (December 25th).

equal for all regions, as shown in the boxplots.

Summer. In the first period of the summer (05-18 to 06-07) we notice the same trend as it was at the end of the 1st lockdown. Then in the centre of the summer the correlations are generally weaker than before, due to the fact that we see a flat behaviour both for mobility and prevalence. However the variation among regions is huge: the blue boxplots in Fig. 6 and 7 in the central part are very wide. For example in the period from 08-26 to 09-15 we see a strong negative correlation between residential mobility and prevalence in the north regions, while a positive one in Sicily and Sardinia. The two regions in this period show a strong negative correlation with transit, as the map cleary states (Fig. 11. For Sardinia for example one can see a rapid increasing for the prevalence on August 23 (Fig. 4), while the transit rapidly decreases. In the end of the summer we come back to a situation we saw at the beginning.

Moreover we observe in Fig. 6 and Fig. 7 that, on average, one is the opposite of another. This can be clearly seen also respectively in the maps (see Fig. 8 and 10 for 1st lockdown and Fig. 9 and 11 for Summer period) showing residential and transit data for every phase are one the opposite of each other. However, usually residential mobility and prevalence show positive τ correlations, while transit mobility and prevalence show negative τ correlations. A general trend that is observed is that the correlation for periods different than summer holidays is well defined, and either very high or very low for all regions, depending on the mobility we are considering.

2nd lockdown. With particular regards to the second lockdown, where regions were grouped according to the restrictions going on, one can see that Kendall's τ correlation computed using residential mobility has still the opposite behavior with respect to the one computed using transit mobility as before. This is true despite the color of the region, see Fig. 12.
Moreover, grouping regions by mobility restrictions and for instance considering red regions we do not observe a clear pattern among them. E.g. for Tuscany, which is red for the 1st phase (11-03 to 12-04), one can see a slightly positive correlation between transit and prevalence. From the plot in Fig. 2 we see that the transit mobility is decreasing in that time and the shifted prevalence too. While for example for Puglia, that is red as well in that period, the aforementioned correlation is negative.
By looking at the boxplot in Fig. 12, we observe that is not possible to find any one-to-one correspondence to the type of mobility restrictions and τ correlations. Indeed, one should compare manually and singularly the graphs

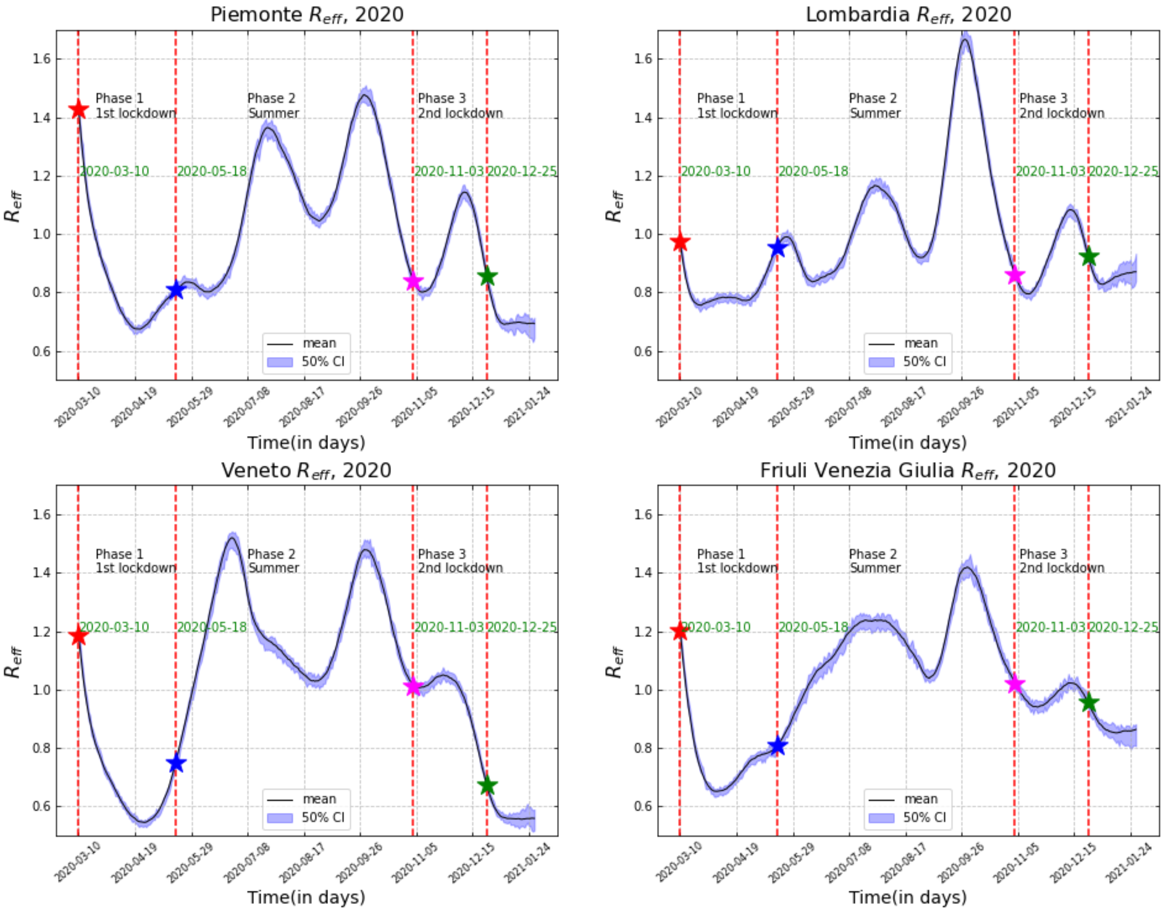


FIG. 14: Plot of R_{eff} vs time(in days) for 4 regions of Italy, marked by red dotted vertical line representing phases and red star → phase 1 start, blue star → end of phase 1 or start of phase 2, magenta star → end of phase 2 or start of phase 3 and finally green star → end of phase 3.

with prevalence and mobility data for each region taken singularly, in order to successfully interpret the maps in 12, as we did as an example in the previous paragraph.

R_{eff} **vs Transit mobility:**

One of the aims of the project is to understand how variations in transit mobility are related to variations in epidemic spreading potential such as R_{eff}. By taking inspiration from the work of Aguilar *et al*[8], Fig. 16 and 15 have been visualized. This procedure was done for all Italian regions, but here only four have been arbitrarily chosen to be presented: other plots are visible in the Appendix. The Loess non-parametric fit was performed on the scatter plot. Each of the three colored markers representing the three phases 1st lockdown, Summer and 2nd lockdown with the increasing intensity of the color adding temporal dimension of increasing time to the scatter plot. We recall that the later a point in time, the darker the intensity in color. From all the regions in Fig. 16, we see a general trend that at the start of the 1st lockdown the R_{eff} is quite high, but in the end of the first lockdown or start of the 2nd lockdown represented by darkest red square marker or lightest triangular marker R value becomes lower as well as mobility is reduced. During the summer, as the time passes we see there is increase in mobility as well as increase in R_{eff}. But however, in the end part of summer both transit mobility as well R_{eff} decreases a little. But then after the start of the second lockdown, we see that transit mobility is relatively low with respect to mid-summer mobility, but as time passes, the mobility as well as R_{eff} increases a little and not by any huge margin.

From the Loess fits, we observe a clear trend from all the plots of Fig. 16 that with the increasing mobility the Reproduction number also increases in some non linear

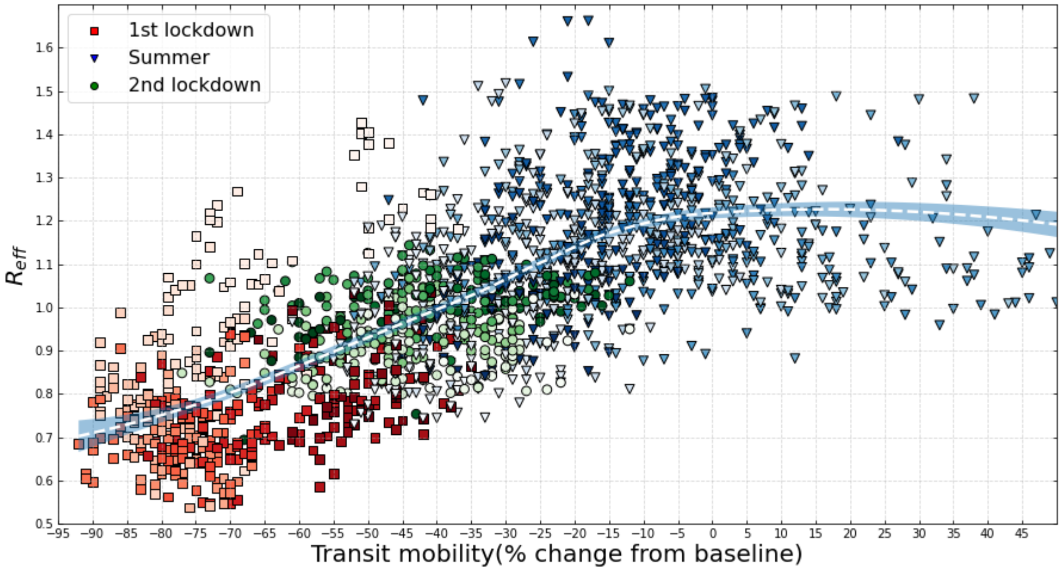


FIG. 15: Plot of R_{eff} vs transit mobility for all regions of Italy, with marker shapes characterizing the phase and the increasing intensity of marker color characterizing the start to end time of each phase. Loess fit with 95% CI is also represented to better observe the trend.

way.

Even the Fig. 15 consisting of all the regions together conveys the same trend as observed from Fig. 16. In the 1st lockdown as time passes the mobility initially decreases along with R_{eff}, however, by the end of 1st lockdown we again observe increase in mobility along with increase in R_{eff}. For the summer, we see a high mobility as expected since people tend to move around a lot since lockdown was lifted as well as it was a holiday season. However in this particular period, we see an increase and then decrease of R_{eff} which can also be observed from Fig. 13. At the start of the 2nd lockdown, since the restriction were stricter as compared to summer period, we observe the white circles at around -55%, but then with time the kind of restrictions were implemented zone wise, and so as a general trend we expect to see increase in mobility along with increase in R_{eff} and we observe the same.

CONCLUSIONS

In our work we tried to spot some correlations between residential and transit mobilities with prevalence by the means of Kendall's τ correlation. We saw that really similar results between regions were returned during 1st lockdown, start of summer and end of summer period as shown in choropleth maps 8,10,9,11. Moreover, we noted as computing the Kendall's τ correlation between prevalence and either residential or transit mobility return essentially opposite results.

For every region $R_{eff}(t)$ was computed by the mean of a MCMC, then relating instead transit mobility evolution with this $R_{eff}(t)$ we noted a clear trend common to all regions. Generally when the transit mobility increases also the R_{eff} increases. Further works may involve to perform the same analysis but letting incidence and prevalence exchange the role, and see whether results can be comparable. One may want to check also whether the lag to be applied to return the largest correlation is constant with respect to the time passing and the type of restrictions applied. In addition, one may want to use also different types of mobility, beside the transit and residential one: Google provides many more types, but different data can be scraped also from other sources. Finally, one should recall that mobility data provided are not representative of the behavior of the whole population.

Another limit for this work was that, due to lack of time, it was not possible to visually and compare curves for all the single regions in Italy thus doing a fine-comparison between epidemiological curves and mobility curves. Indeed our analysis, despite we wanted to achieve results as general as possible, has to be tuned for every region: specially during summer as an example we expect huge variations in transit mobility to touristic areas, rather than rural. Moreover during the 2nd lockdown no

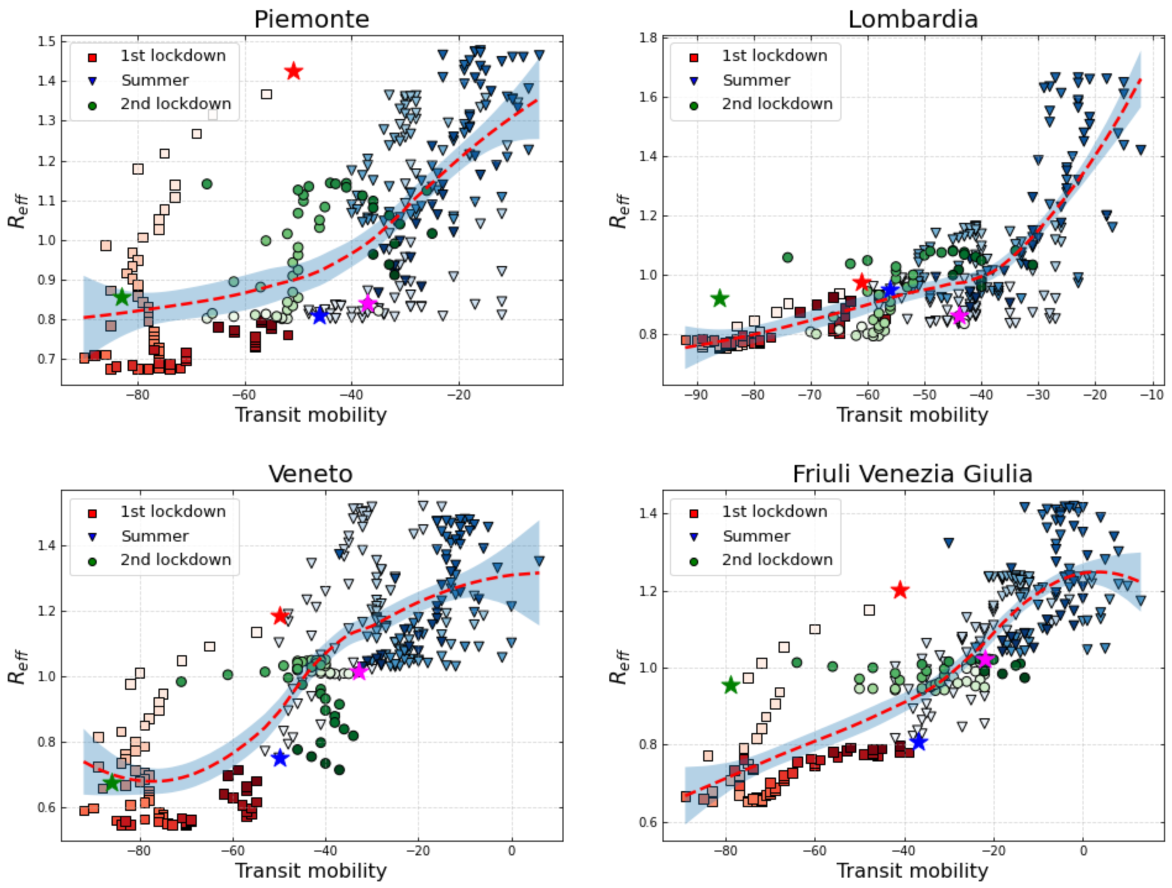


FIG. 16: Plot of R_{eff} vs transit mobility for four regions of Italy, with marker shapes characterizing the phase and the increasing intensity of marker color characterizing the start to end time of each phase. Loess fit with 95% CI is also represented to better observe the trend. Red star → phase 1 start, blue star → end of phase1 or start of phase 2, magenta star → end of phase 2 or start of phase 3 and finally green star → end of phase 3.

pattern that characterizes the regions of the same colour is found. Instead, if looking at single region behaviours, the correlation strongly depended on the actual epidemiological situation at that moment: regions sharing same mobility restrictions had very different Kendall's τ.

To conclude, mobility and prevalence are correlated and we can say that mobility can be a proxy of human encounters which is the cause of increase in infections. Variation of R_{eff} is found to be related to variation in the mobility, i.e. increase in mobility leads to increase in Effective Reproduction rate.

The code exploited for this report is available at the link in references [10].

[1] Covid-19 italia - monitoraggio situazione. Available at `https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv`.

[2] *DPCM 2020 April 10th.* Available at `http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioNotizieNuovoCoronavirus.jsp?lingua=italiano&menu=notizie&p=dalministero&id=4447`.

[3] *DPCM 2020 April 26th.* Available at `http://www.governo.it/sites/new.governo.it/files/Dpcm_img_20200426.pdf`.

[4] *Google LLC "Google COVID-19 Community Mobility Reports".* Available at `https://www.google.com/covid19/mobility/`.

[5] *Italy DPCM December 4.* http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioNotizieNuovoCoronavirus.jsp?lingua=english&id=5211.

[6] *Italy DPCM region restrictions for 2nd lockdown.* https://www.ictp.it/ictp_covidresponse/italian-government-actions.aspx.

[7] The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020. PMID: 32150748.

[8] Javier Aguilar, Aleix Bassolas, Gourab Ghoshal, Surendra Hazarie, Alec Kirkley, Mattia Mazzoli, Sandro Meloni, Sayat Mimar, Vincenzo Nicosia, Jose J. Ramasco, and Adam Sadilek. Impact of urban structure on covid-19 spread, 2020.

[9] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018.

[10] Camilla Quaglia Andrea Nicolai, Karan Kabbur Hanumanthappa Manjunatha. Life data epidemiology project. https://github.com/andrybicio/LifeDataEpidemiology_Project, 2021.

[11] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, Mar 2018.

[12] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David S.C. Hui, Bin Du, Lan-juan Li, Guang Zeng, Kwok-Yung Yuen, Ru-chong Chen, Chun-li Tang, Tao Wang, Ping-yan Chen, Jie Xiang, Shi-yue Li, Jin-lin Wang, Zi-jing Liang, Yi-xiang Peng, Li Wei, Yong Liu, Ya-hua Hu, Peng Peng, Jian-ming Wang, Ji-yang Liu, Zhong Chen, Gang Li, Zhi-jian Zheng, Shao-qin Qiu, Jie Luo, Chang-jiang Ye, Shao-yong Zhu, and Nan-shan Zhong. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, 382(18):1708–1720, 2020.

[13] Guidotti and Ardia. Covid-19 data hub. Available at https://covid19datahub.io/articles/iso/ITA.html, 2020.

[14] William Jacoby. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19:577–613, 12 2000.

[15] Maurice G. Kendall. *The treatment of ties in ranking problems.* Biometrika Vol. 33, No. 3, pp. 239-251, 1945.

[16] Naomi Krzywinski M., Altman. Visualizing samples with box plots. *Nature Methods*, 11, 2014.

[17] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah Meredith, Andrew S. Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020. PMID: 32150748.

[18] Linton N. M. Akhmetzhanov A. R. Nishiura, H. Serial interval of novel coronavirus (covid-19) infections. *International journal of infectious diseases*, 93:284–286, 2020.

[19] Spada Paolo. Covid-19 italia — i grafici e le mappe interattive di pillole di ottimismo. https://public.flourish.studio/story/722265/?fbclid=IwAR2Z5En_B7Bg6fLCubR6sPgMV7P8tTRhpd6Mv7kB__Zai3YxgMc5RYS-syQ, 2020.

[20] Fonnesbeck C. Salvatier J., Wiecki T.V. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2(e55):577–582, 2016. PMID: 32150748.

[21] M. Sulyok and M. Walker. Community movement and covid-19: a global study using google's community mobility reports. *Epidemiology and Infection*, 148:e284, 2020.

[22] Kevin Systrom, Thomas Vladek, and Mike Krieger. Project title. https://github.com/rtcovidlive/covid-model, 2020.

[23] Bo et al. Xu. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific Data*, 7, 2020.

APPENDIX

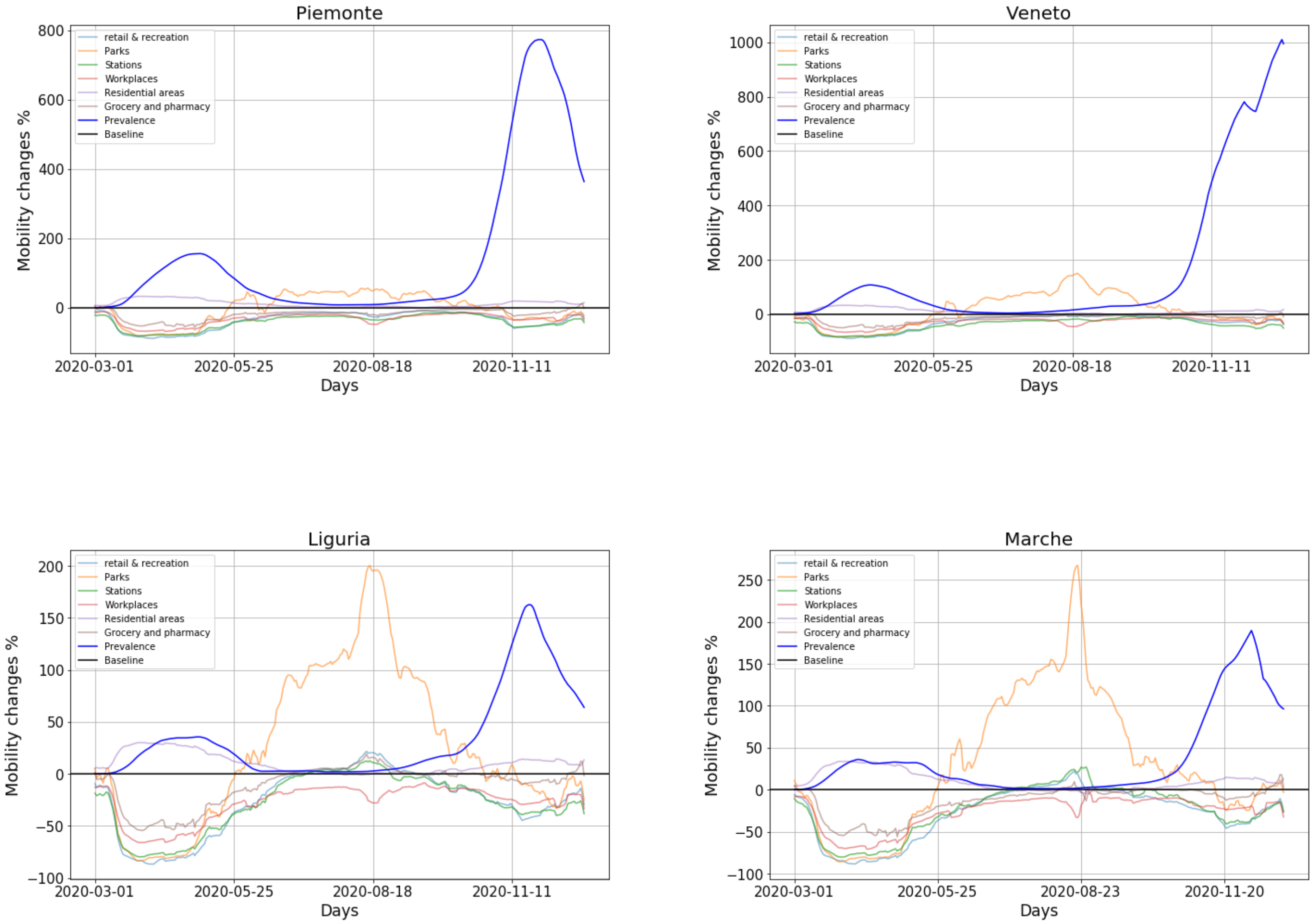


FIG. A.1: Behaviour of the prevalence curve compared with Google's mobility. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility

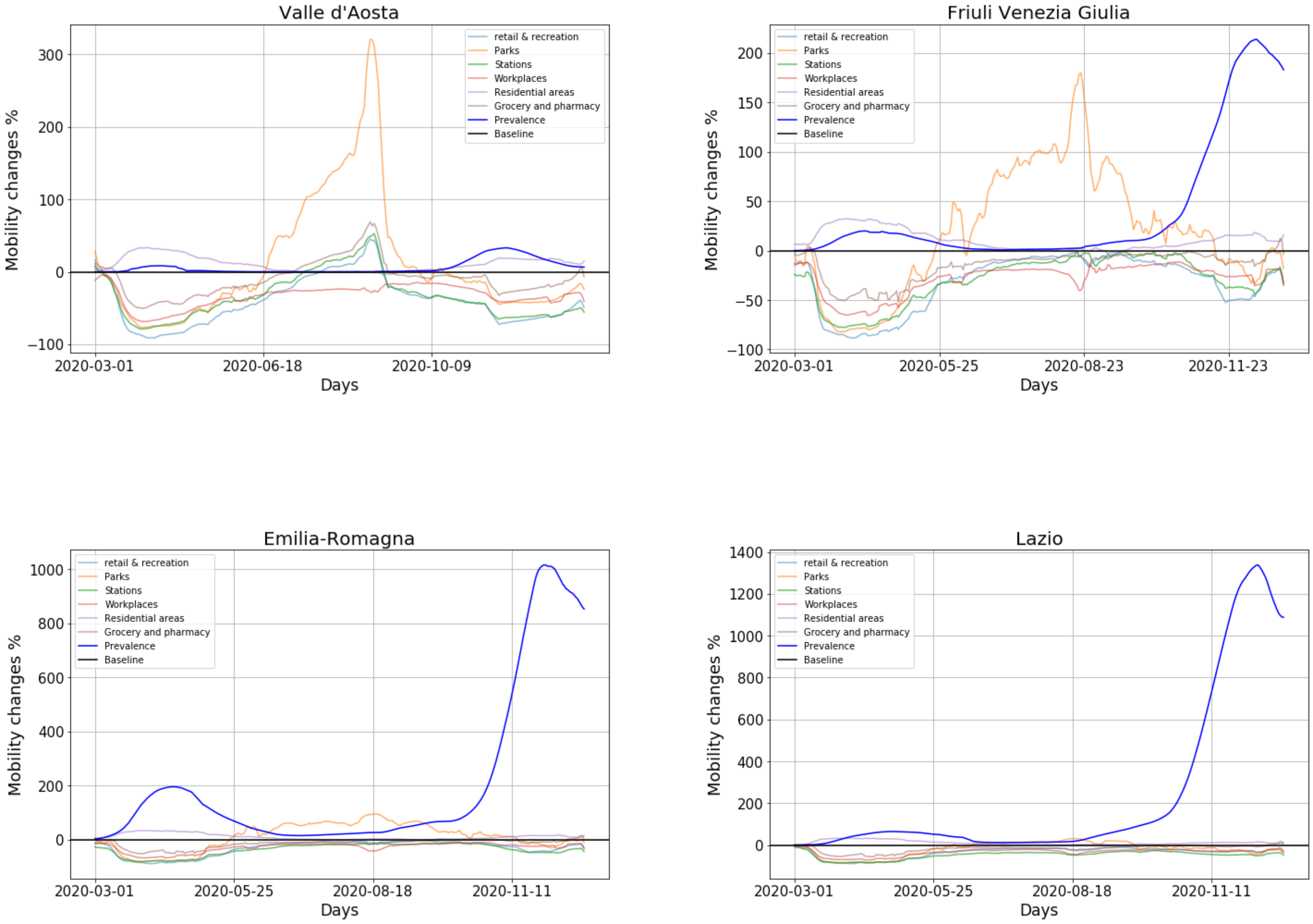
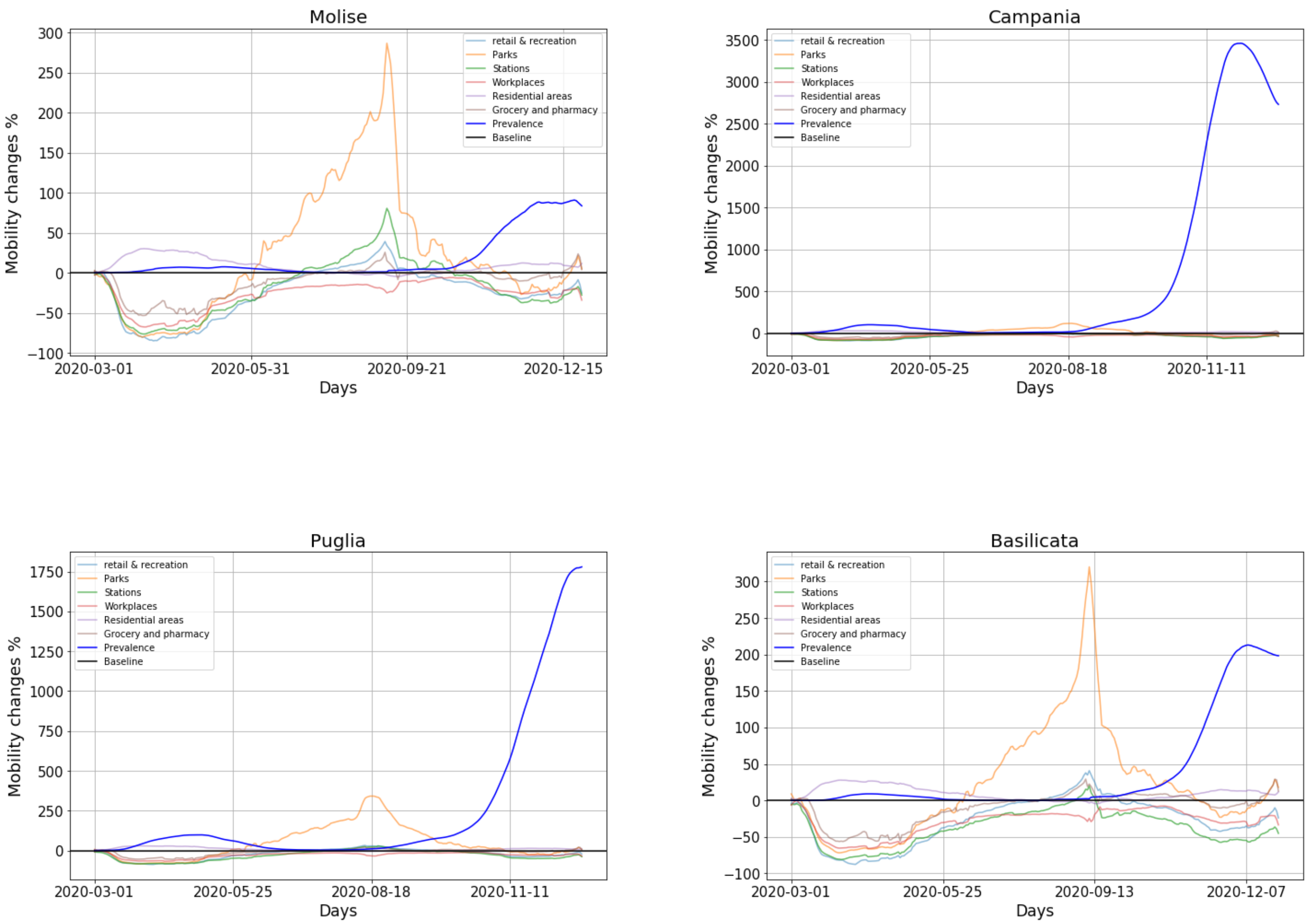


FIG. A.2: Behaviour of the prevalence curve compared with Google's mobility data. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility



Molise

Campania

Puglia

Basilicata

FIG. A.3: Behaviour of the prevalence curve compared with Google's mobility data. The period displayed is from 2020-03-01 to 2020-12-25 and to both data a rolling average with 7 days window was applied. Prevalence is rescaled of the same factor in both graphs in order to compare it with mobility

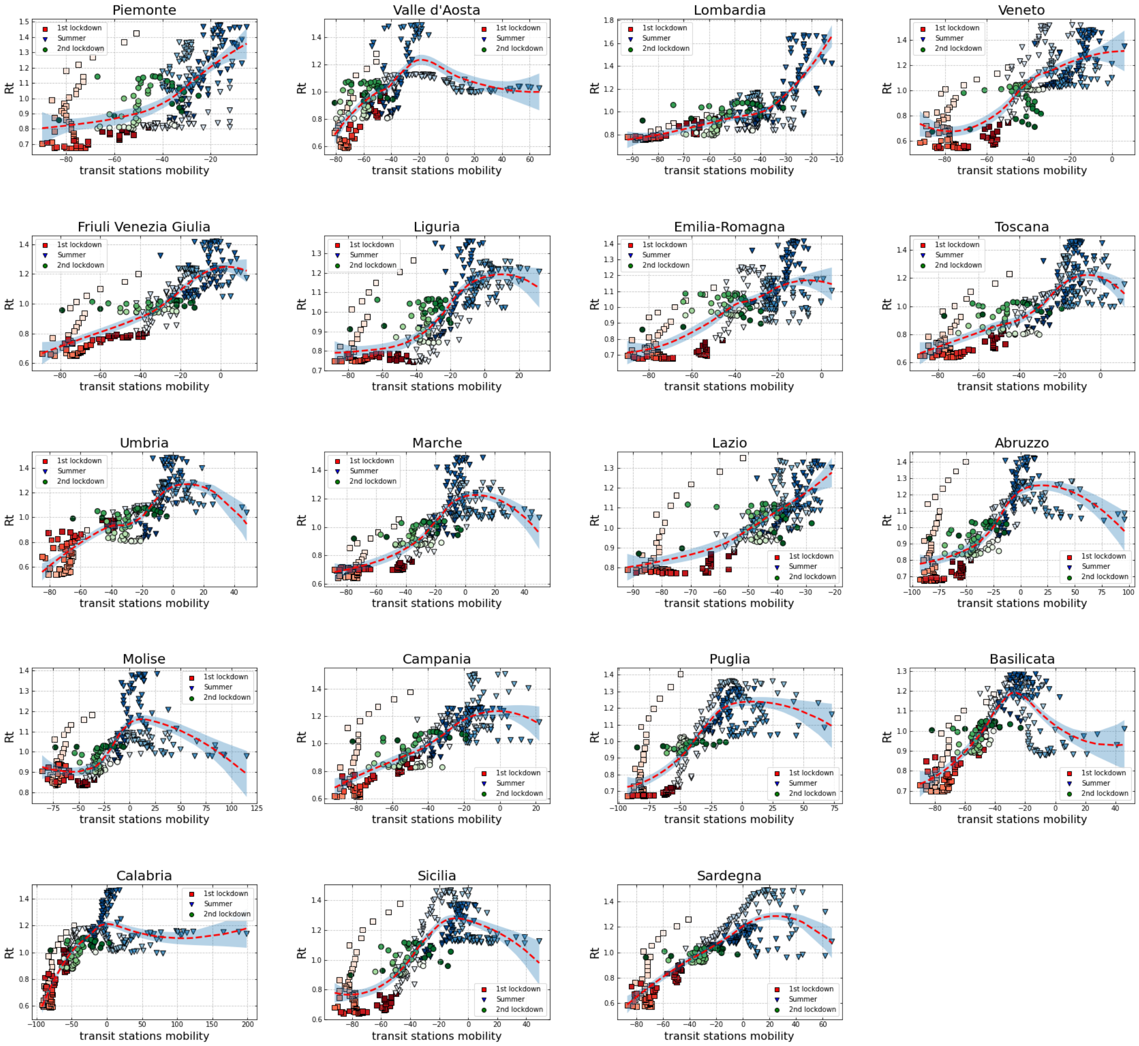
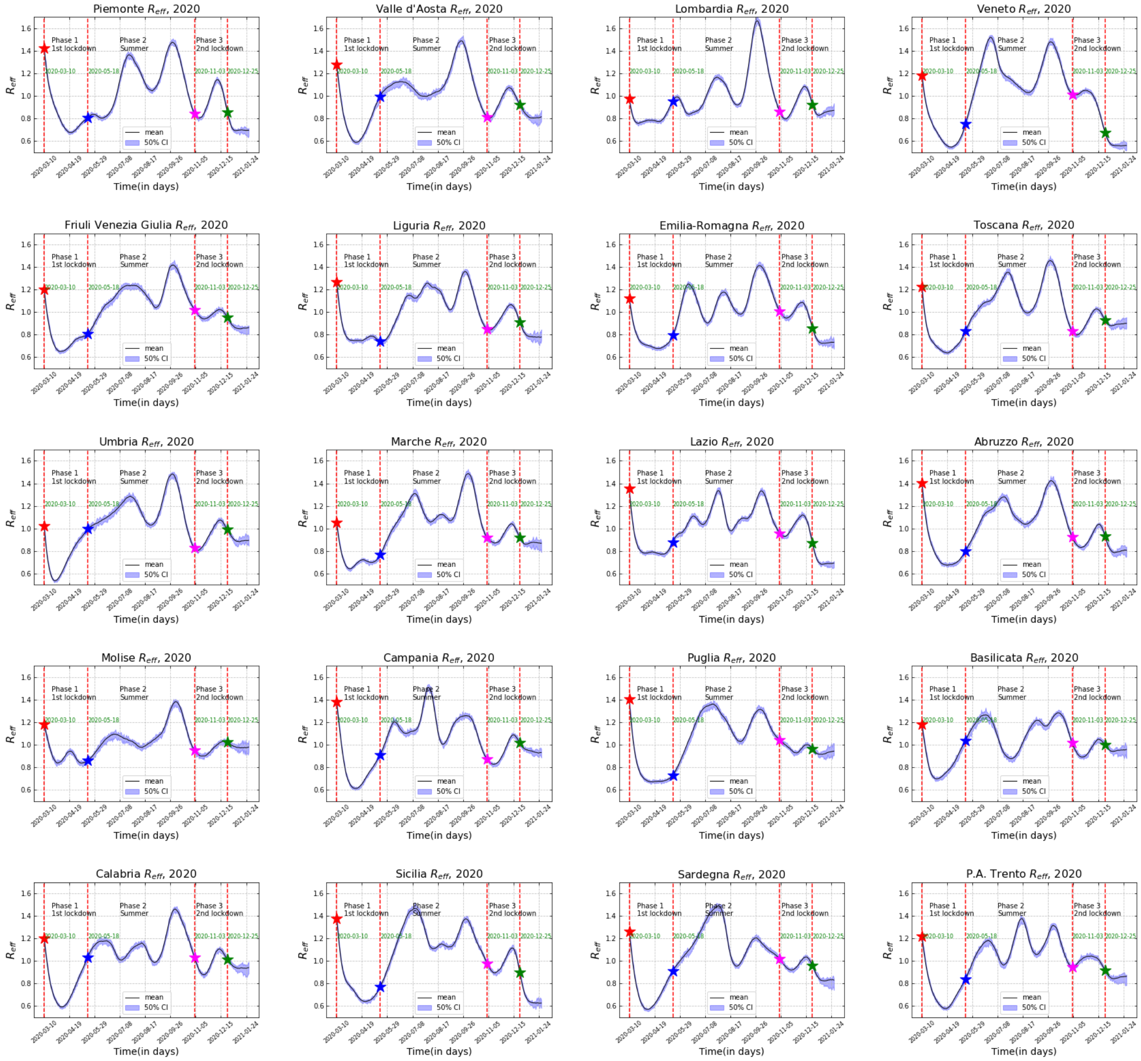


FIG. A.4: Plot of R_{eff} vs transit mobility for each region of Italy, with marker shapes characterizing the phase and the increasing intensity of marker color characterizing the start to end time of each phase. Loess fit with 95% CI is also represented to better observe the trend.



Piemonte R_{eff}, 2020

Valle d'Aosta R_{eff}, 2020

Lombardia R_{eff}, 2020

Veneto R_{eff}, 2020

Friuli Venezia Giulia R_{eff}, 2020

Liguria R_{eff}, 2020

Emilia-Romagna R_{eff}, 2020

Toscana R_{eff}, 2020

Umbria R_{eff}, 2020

Marche R_{eff}, 2020

Lazio R_{eff}, 2020

Abruzzo R_{eff}, 2020

Molise R_{eff}, 2020

Campania R_{eff}, 2020

Puglia R_{eff}, 2020

Basilicata R_{eff}, 2020

Calabria R_{eff}, 2020

Sicilia R_{eff}, 2020

Sardegna R_{eff}, 2020

P.A. Trento R_{eff}, 2020

FIG. A.5: Plot of R_{eff} vs time(in days) for each region of Italy, marked by red dotted vertical line representing phases and red star → phase 1 start, blue star → end of phase1 or start of phase 2, magenta star → end of phase 2 or start of phase 3 and finally green star → end of phase 3.