

# Emergent circulation patterns from anonymized mobility data: Clustering Italy in the time of Covid

Jules Morand,<sup>1,\*</sup> Shoichi Yip,<sup>1</sup> Yannis Velegarakis,<sup>2</sup> Gianluca Lattanzi,<sup>1</sup> Raffaello Potestio,<sup>1</sup> and Luca Tubiana<sup>1,†</sup>

<sup>1</sup>*Physics Department, University of Trento, via Sommarive, 14 I-38123 Trento, Italy,  
INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy*

<sup>2</sup>*Information and Computing Science, University of Trento, Italy and Utrecht University, Netherlands  
(Dated: June 9, 2023)*

Using anonymized mobility data from Facebook users and publicly available information on the Italian population, we model the circulation of people in Italy before and during the early phase of the SARS-CoV-2 pandemic (COVID-19). We perform a spatial and temporal clustering of the movement network at the level of fluxes across provinces on a daily basis. The resulting partition in time successfully identifies the first two lockdowns without any prior information. Similarly, the spatial clustering returns 11 to 23 clusters depending on the period (“standard” mobility *vs.* lockdown) using the greedy modularity communities clustering method, and 16 to 30 clusters using the critical variable selection method. Fascinatingly, the spatial clusters obtained with both methods are strongly reminiscent of the 11 regions into which emperor Augustus had divided Italy according to Pliny the Elder. This work introduces and validates a data analysis pipeline that enables us: i) to assess the reliability of data obtained from a partial and potentially biased sample of the population in performing estimates of population mobility nationwide; ii) to identify areas of a Country with well-defined mobility patterns, and iii) to distinguish different patterns from one another, resolve them in time and find their optimal spatial extent. The proposed method is generic and can be applied to other countries, with different geographical scales, and also to similar networks (e.g. biological networks). The results can thus represent a relevant step forward in the development of methods and strategies for the containment of future epidemic phenomena.

In order to minimize the impact of epidemics such as the recent SARS-CoV-2 pandemic [1] on society, governments must take far-reaching decisions that considerably affect the lives of their citizens. Some common measures deployed during the pandemic were the adoption of personal protection devices such as face masks [2, 3], contact tracing aimed at identifying and confining infectious subjects [4–9], and the use of various forms of lockdown to dampen large-scale contagion [10–15].

Italy was the first European country to impose a national lockdown and has seen the implementation of three nationwide lockdowns: between March and April 2020, in January 2021, and in April 2021. Detailed studies have been carried out on the initial propagation of the epidemic in Italy [16, 17], on the first confinement [18], and the relaxation of the latter [19], discussing the necessity and the implementation of such restrictive measures.

While lockdowns are certainly effective in curbing the rise of infections, their imposition severely affects the life and health of citizens [20–22]. The extent of their deployment needs to be optimized both in space and time to minimize the number of people affected while guaranteeing the safety of the population. For this reason, after the first phase of the pandemic, the Italian government delegated part of the responsibility of restrictions to regional governments, which were forced to curb the movements of their citizens whenever the effective reproduction number  $R_t$  (i.e. the average number of new in-

fections caused by a single infected individual at time  $t$ ) went above 1 [23–25]. Imposing regional lockdowns instead of national ones is a sensible strategy. However, it is not guaranteed that existing administrative regions correspond to the best subdivisions of a state to control the spread of epidemics.

In general, a diffusion process in human society depends on the complex structure of the underlying network of interactions. At the individual scale, several studies make use of social experiments in recording the contacts of a group of people *via* special devices; this was done e.g. in a summer camp for children in Italy [26] or with primary and high-school students in France [27–29]. Such data can then be used to generate a time-dependent network of contacts that can be later used to simulate the diffusion of an epidemic and see how it develops at the scale of the single individual [30, 31]. At larger scales, privacy concerns and pragmatic necessities can make it preferable to turn towards the usage of meta-population network models [32, 33]. This can be done for example at a national [34] or international level [35, 36], or at multiple levels through the usage of multiscale information on mobility [37]. These and other studies can be informed by anonymized data such as airplane traffic [35, 36] or social network location data [38].

Here, by analyzing the mobility of the Italian population in the period between January 2020 and May 2022, we show how a data-driven meta-population approach can be used to identify the optimal spatial subdivision of a state to control an epidemics, as well as to verify a posteriori the effectiveness of lockdowns. To do so, we first estimate the mobility of the Italian population

---

\* jules.morand@unitn.it

† luca.tubiana@unitn.it

at the level of provinces (small administrative regions between municipalities and regions) thanks to Facebook (FB) data obtained through META's *Data for Good* program [39]. To check the reliability of these data, we compute the population density vector (i.e. the normalized vector of relative populations in the Italian provinces) obtained from META's data against the one derived from the independent data set of the Italian National Institute of Statistics (Istat [40]), containing the official projected census for January 1st, 2020 [41]. The good agreement between these vectors shows that the data collected by META through the FB geolocalisation service provide a good estimate of the distribution of the Italian population. Then, we show how a clustering in time is able to correctly identify the first two national lockdowns, which were strictly enforced by the Italian state. Finally, we use the most characteristic mobility matrices for confined and non-confined phases to find the optimal spatial clustering of Italian provinces. To do so, we employ two different clustering methods that partition Italy into clusters of provinces matching with areas having cultural, social, and commercial affinities during 'ordinary' times, and into smaller clusters during confinement times. Applications of this approach to other Countries, scales, and other complex networks are discussed.

## I. RESULTS

Our approach to characterize the behavior of the Italian population is based on movement data between provinces. These are administrative entities in between regions and municipalities, usually containing between one and three hundred thousand people, with those corresponding to major cities such as Rome, Naples, Milan, Turin, and Palermo having more than a million inhabitants [40].

As explained in detail in the Methods section, we consider 106 provinces and extrapolate the movement of their respective populations from FB users' data provided by META's data for good program [39]. The dataset we used provides the number of FB users in each province  $i$ ,  $n_i$ , as well as the number of users moving between two provinces (or within a province),  $n_{ij}(t)$ , every 8 hours in the period between January 2020 and May 2022.

### A. Transition matrices

The data from META allow us to compute the 8-hours transition rate between two provinces  $i$  and  $j$ , defined as follows:

$$\Pi_{ij}(t) = \frac{n_{ij}(t)}{\sum_j n_{ij}(t)}. \quad (1)$$

Note that the denominator ensures that, for every province  $i$ ,  $\sum_j \Pi_{ij} = 1$ , thereby guaranteeing that  $\Pi$  can be used as a stochastic matrix.

To get an idea of what the data look like, the time evolution of one link  $\Pi_{ij}$ , reporting the mobility from the province of Agrigento ( $i = AG$ ) to that of Caltanissetta ( $j = CL$ ), is plotted in Fig. 1a). Daily averaged values are reported in blue, weekly averaged ones in red, and the corresponding entry in the mean matrix of Eq. 6 (see materials and methods) in a black dashed line. The lockdown periods are indicated by grey-shaded vertical bars. Seasonal effects are clearly visible from the comparison of the daily data and the corresponding weekly averaged ones.

To remove seasonal fluctuations in  $\Pi$  (day vs night, weekdays vs weekends) we redefine  $\Pi$  as the daily transition rate between provinces averaged over the three days before and three days after, see Materials and Methods section. Finally, it is convenient to consider the mean transition matrix over the whole period,  $\bar{\Pi}$ . The directed graph associated with  $\bar{\Pi}$  is displayed in Fig. 1b).

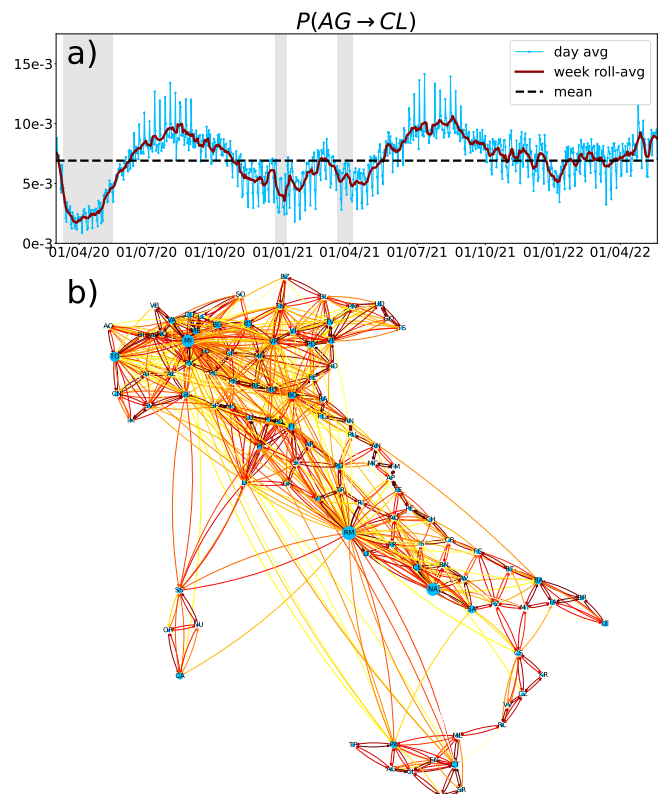


FIG. 1. a) AG→CL (Agrigento to Caltanissetta provinces) link vs time. Daily average probability (blue) and 7-day rolled-average probability (red), and overall probability averaged in time (black dashed line). b) Representation of the directed graph defined by the Matrix  $\bar{\Pi}$  (Eq.6). Arrows represent the mean probability links,  $\bar{\Pi}_{ij}$ , between Italian provinces  $i$  and  $j$ , and are scaled in size and color according to the value of the link (from light yellow to dark red). The size of the nodes is proportional to the population (vector  $\rho^*$  of Fig. 2). Self-links  $\bar{\Pi}_{ii}$  are not shown as they have a much bigger value ( $\sim \times 10^2$ ) than the non-diagonal links, as most people do not move out of their province.

## B. Homogeneity and representativeness of FB data

We assume that the FB users in the database are homogeneously distributed across provinces, and move in a manner that is on average similar to that of the rest of the population. To validate these assumptions we proceed as follows.

First, we monitor the fraction of FB users over the total population of the province according to Istat; this ratio is defined as  $\bar{n}_i/n_i^{\text{Istat}}$ , where  $\bar{n}_i = \langle n_i^h \rangle$  is the number of FB users in province  $i$  averaged over the whole time series. The results, reported in Fig. 2a), show that in all provinces this fraction remains between 3% and 7%, and that FB users are roughly homogeneously distributed across the country (Fig.S2 of supporting information displays the vectors with all provinces two-letter codes).

A more quantitative validation of both assumptions can be obtained by considering the *population density vectors* obtained both from the official census of Istat in 2020 and from FB users' data. These are defined as follows:

$$\boldsymbol{\rho} = \left( \frac{n_1}{n_{tot}}, \dots, \frac{n_N}{n_{tot}} \right)^T \quad (2)$$

where  $n_1, \dots, n_N$  are the populations of the  $N$  provinces, and  $n_{tot} = \sum_{i=1}^N n_i$  is the total population. The populations  $n_i$  can be obtained from either:

- Istat data,  $\boldsymbol{\rho}^{\text{Istat}}$ ,
- the FB population dataset,  $\boldsymbol{\rho}^{\text{FB}}$ ,

The above normalization, Eq. 2, sets  $|\boldsymbol{\rho}| = 1$  and allows us to compare the different vectors. In addition, our approach makes it possible to compare another population density vector,  $\boldsymbol{\rho}^*$ , obtained from the mean transition matrix  $\bar{\Pi}$  extracted from the FB movement dataset.

In the graph described by  $\bar{\Pi}$  there is a non-zero probability to reach any node from any other one in a finite number of steps, that is, the graph is strongly connected and aperiodic, and random walks over it are ergodic. The Perron-Frobenius theorem then ensures that  $\bar{\Pi}$  has a non-degenerate highest eigenvalue. With our normalisation of  $\bar{\Pi}$  this is  $\lambda^* = 1$ , and its associated left eigenvalue  $\rho^*$  is the only stationary state of the system, satisfying:

$$\rho_i^* \bar{\Pi}_{ij} = \bar{\Pi}_{ji} \rho_j^*.$$

Therefore, any non-trivial distribution vector over the nodes of our network will converge to  $\boldsymbol{\rho}^*$  after a sufficiently long time (see supporting information: section 3)

If the movements described by  $\bar{\Pi}$  are consistent with the Istat population data, the stationary density vector  $\boldsymbol{\rho}^*$  must be in good agreement with the Istat density vector  $\boldsymbol{\rho}^{\text{Istat}}$ . This is indeed the case, as shown in Fig. 2b) and c).

Fig. 2, panel b) displays the population density vectors  $\boldsymbol{\rho}^{\text{FB}}$  and  $\boldsymbol{\rho}^*$ , on a log-log scale against  $\boldsymbol{\rho}^{\text{Istat}}$ . The

provinces are sorted from least to most populated according to Istat data. We see a good agreement within the FB data themselves, which is also a benchmark of our extraction and preparation of the data.

Moreover, the standard deviations of  $\boldsymbol{\rho}^{\text{FB}}$  and  $\boldsymbol{\rho}^*$  from the Istat vector (panel c of Fig. 2) are in very good quantitative agreement with the Istat data. However, we notice that the most populated provinces, Rome, Milan, Naples, Turin, (RM, MI, NA, TO) are slightly overestimated and that the less populated provinces are slightly underestimated especially by the  $\boldsymbol{\rho}^*$  vector. This can be explained by the fact that all links with less than 10 people are ignored for privacy reasons.

## C. Transition matrices time series

Having validated the FB data, we proceed to extract the information contained in the time series of weekly-averaged daily transition matrices. First of all, we notice that diagonal elements  $\Pi_{ii} \geq 0.9$ , meaning that most movements happen within provinces. Second, and most notably, we find that while the time series of the probability to move between different provinces can vary by an order of magnitude, as shown in Fig. 3a), the movement pattern of single provinces can be brought to collapse on two master curves with an appropriate rescaling, see Fig. 3b)-e). Specifically, this can be done by considering the normalized probability to move out of a province,

$$P_{out}^i(t)/\overline{P_{out}} = (1 - \Pi_{ii}(t))/(1 - \bar{\Pi}_{ii}), \quad (3)$$

shown in Fig. 3b), c) and the normalized probability to move into a province,

$$P_{in}^i(t)/\overline{P_{in}} = \sum_{j \neq i} \Pi_{ji}(t) / \sum_{j \neq i} \bar{\Pi}_{ji}, \quad (4)$$

reported in Fig. 3d), e). As can be seen from panels c), e) all provinces display a similar behavior in these two quantities, and the first two lockdowns become apparent as periods of low mobility. The Z-score, i.e. the time average of the fluctuation of  $P_{in}^i(t)$  and  $P_{out}^i(t)$  with respect to the mean over provinces, is defined and displayed in supporting information (Fig.S6).

Interestingly, we also note that some provinces show a large deviation in both quantities in correspondence of summer and winter months. To rationalize this, we look at the provinces showing peaks of mobility in those periods, and found them to correspond with those having a high touristic vocation, as for example Belluno (BL) and Trento (TN), near the Dolomites, and Sicilian provinces, see Fig. 3b), d). While at first the fact that Rome and Venice (VE) do not show these peaks might be unexpected, we recall that our data only follow the movement of Italian citizens, and that during Covid there was a strong push to take holidays outside of cities.

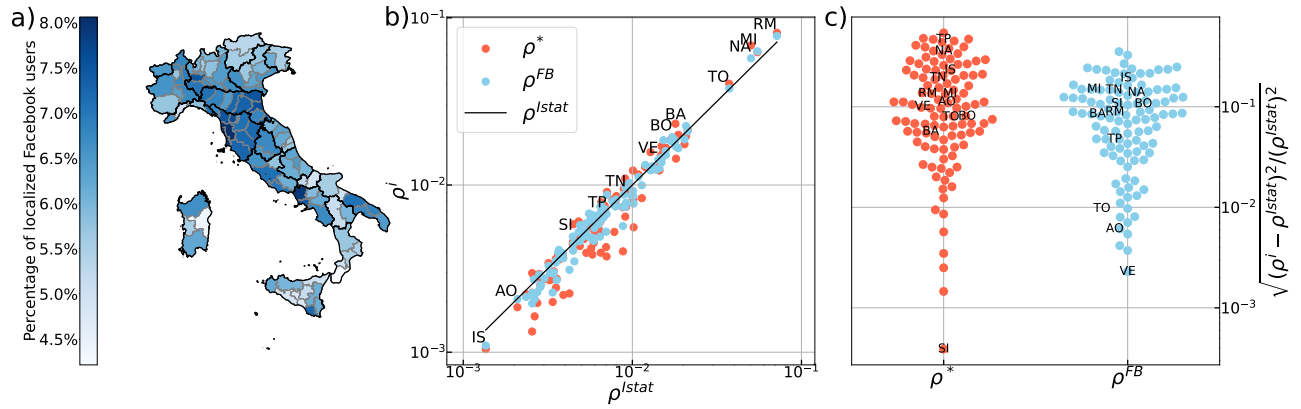


FIG. 2. a) Fraction of FB users that have shared their location over the official province population obtained from the Istat 2020 census,  $\bar{n}_i/n_i^{\text{Istat}}$ , for each province  $i$ . b) Comparison of the different population density vectors from FB and Istat data:  $\rho^{\text{FB}}$  and  $\rho^*$  are plotted against  $\rho^{\text{Istat}}$ . c) Standard deviation of the vectors  $\rho^{\text{FB}}$  and  $\rho^*$  from the  $\rho^{\text{Istat}}$  vector.

#### D. Temporal Clustering

In order to reach our goal, that is to use movement data to identify spatial communities, we first need to ensure that the information contained in the transition matrices  $\Pi(t)$  is sufficient to identify the national lockdown periods.

To do this, we cluster the daily movement matrices into two groups based on the distance induced by the matrix-matrix scalar product, as described in the Materials and Methods section. The results are reported in the top panel of Fig. 4, where each matrix is represented by the average probability for people to move out of their province at time  $t$ :

$$\langle P_{\text{out}} \rangle (t) = \frac{1}{N} \sum_{i=1}^N 1 - \Pi_{ii}(t) = 1 - \frac{1}{N} \text{Tr}(\Pi(t)). \quad (5)$$

The two temporal clusters  $C_0$  and  $C_1$  are represented by light blue dots and dark red stars, respectively, and the latter clearly identifies the first two national lockdown periods, delimited by the vertical shaded areas. Although the third lockdown period is not identified by the clustering, we argue that this is because it has not been strictly imposed, nor was it effectively respected, as can also be seen from the mobility plots of Fig. 4, and Fig. 3b)-e).

While in order to fully understand the behaviour of the new infections and hospitalization curves, one would need to take into account a number of factors, such as for example population density and temperature variations [42], the repercussions of the confinements on the evolution of the epidemics are clearly visible in Fig. 4: the lockdowns are all followed by a decrease in the number of new cases and new hospitalizations, as expected [12, 13]. Furthermore, we can notice that the curves for  $\langle P_{\text{out}}(t) \rangle$ , new cases, and new hospitalizations, are in general anti-correlated, with mobility decreasing in correspondence to increases in the other two curves, which then reach a peak

and decrease. The reduction in mobility outside of national lockdowns is arguably due to individual decisions and even more to local movement restrictions applied by regions; the fact that a lower mobility leads to a decrease in the number of infections is a standard prediction of epidemic models.

FB data thus entail mobility features that are in agreement with the history of the Italian government's decisions and their repercussions on the population's behaviour, validating their usage in modeling epidemics and social phenomena more in general.

#### E. Optimal Spatial Clustering

We can now perform a spatial clustering of the most representative matrices of the two temporal clusters obtained for the confined and unconfined situations. To this aim, we define for each temporal cluster ( $C_k$ ,  $k = 0, 1$ ):

- the *mean transition matrices*  $\bar{\Pi}^{C_k}$ ,
- the *most representative transition matrices*  $\tilde{\Pi}^{C_k}$ ,
- the *most representative current matrices*  $J^{C_k} = \tilde{\Pi}^{C_k} \rho^{\text{Istat}}$ .

We then use two different methods to perform the clustering itself:

- the Greedy Modularity Communities (GMC) method, which uses  $J^{C_k}$ , i.e. the flux of people moving, and finds the number of clusters that maximizes *modularity*, a concept from graph theory. This algorithm optimizes the clustering such that the inner links of clusters are stronger than the outer ones.
- the critical variable selection (CVS) method, which makes use of  $\tilde{\Pi}^{C_k}$ , i.e. the probability of a sin-

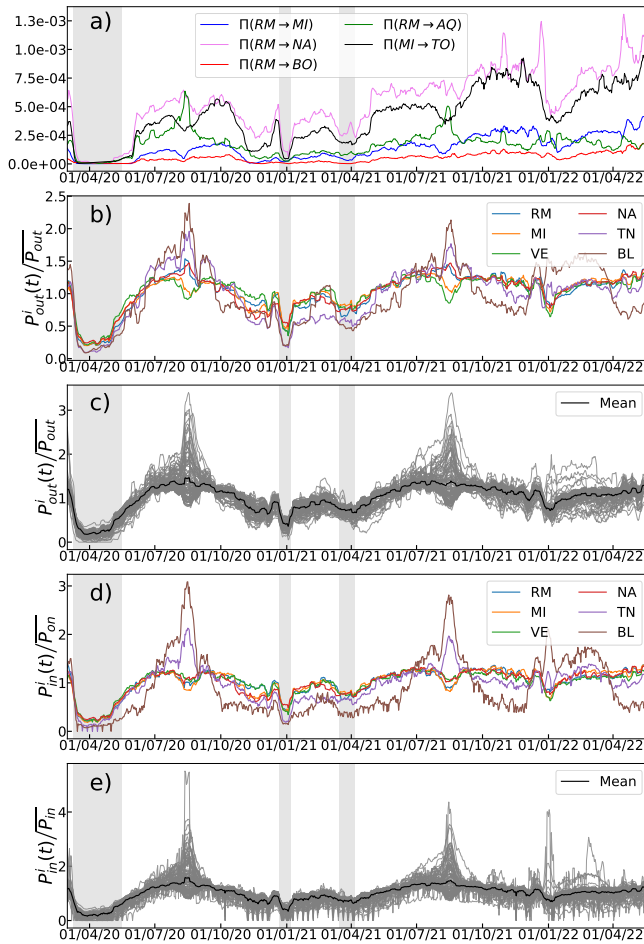


FIG. 3. a) Examples of some representative transition probability links. b) and c) Probability of going out of the province versus time b) for  $i = \text{BL, MI, NA, RM, TN, and VE}$  and c) for the mean over the province in black and the whole distribution of grey. d) and e) Probability of going in the province versus time d) for  $i = \text{BL, MI, NA, RM, TN, and VE}$  and e) for the mean over the province in black and the whole distribution of grey. All probability distributions have been plotted and re-scaled by their temporal average to obtain a collapse of the curves. Gray shaded areas represent national lockdown periods.

gle person to move, and finds the number of clusters that maximizes the *relevance*, a quantity introduced in information theory. This method searches for the clustering that minimizes information loss with respect to a full description of the dataset [43].

The details of both strategies are reported in the Materials and Methods section and graph representation of the most representative matrix in each case can be found Fig.S9 and Fig.S10 of supporting information. We observe here that, although in principle geographically distant provinces could be grouped together (e.g. in the case of highly-connected cities such as Rome, Naples, Milan, and Turin), the clusters found by both methods are

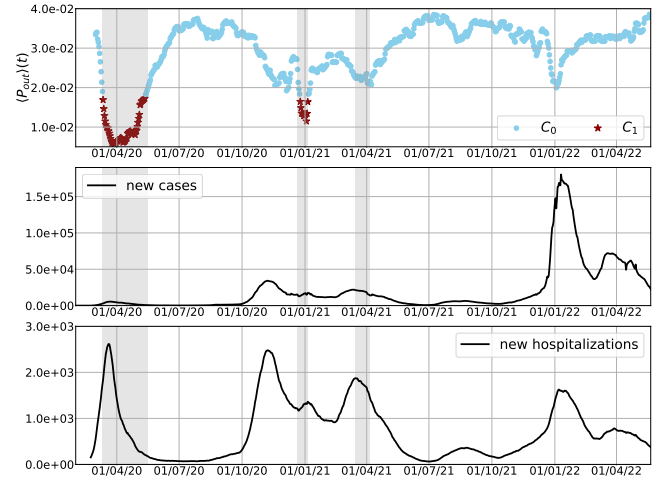


FIG. 4. Top panel: Mean mobility  $\langle P^{out} \rangle(t)$  versus time. The light blue dots and dark red stars illustrate the two temporal clusters of transition matrices series. Gray-shaded areas represent national confinement periods. Center panel: Number of new cases of COVID-19 per day in the whole of Italy versus time. Bottom panel: Number of new hospitalized cases due to COVID-19 in Italy per day.

composed of physically proximal provinces, which can be reached one from the another without having to cross other clusters. This is a non-trivial result, as neither method relies on the notion of geographical distance.

#### Non confined

Fig. 5a,c) represent the clustering of the most representative matrix of the unconfined temporal cluster ( $C_0$  in blue in the top panel of Fig. 4), corresponding to an ‘ordinary’ Italian mobility situation; the top map is obtained employing the greedy modularity method, while the bottom one makes use of the CVS approach.

The two methods return slightly different partitions: for the greedy modularity (top), the Italian provinces are grouped in 11 clusters corresponding to well-defined geographical areas, while 16 groups are found using the CVS scheme. Moreover, apart from a few border cases, the clusterings almost perfectly reproduce well-known cultural and commercial ‘blocks’ within the Country. For example, the green cluster corresponds to the *Triveneto* area (that is Veneto, Friuli-Venezia Giulia, and Trentino-Alto Adige), while Sardinian provinces are fully grouped in their own cluster. The time series of outward and inward probabilities for each province are also displayed in the supporting information (Fig.S7 and Fig.S8) with a highlight for each optimal spatial cluster obtained with the greedy modularity method.

It is interesting to observe that the 11 clusters found through modularity resemble quite strongly those reported by Pliny the Elder [44, 45], according to whom Emperor Augustus divided Italy into 11 regions around

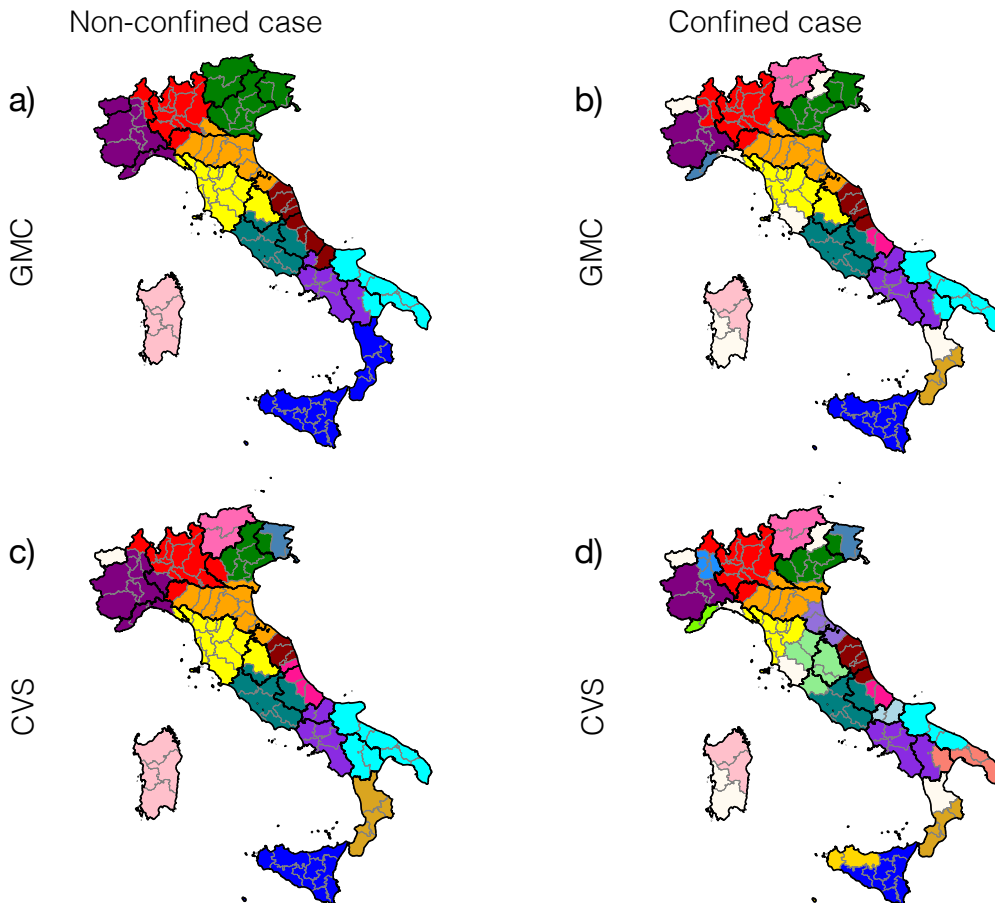


FIG. 5. Spatial community clustering obtained with Greedy Modularity, a), b), and Critical Variable Selection, c), d). Panels a), c) report the communities identified by the two methods during the non-confined periods. Panels b), d), during confinement. Grey lines represent the borders of the provinces while bold black lines delimit administrative regions.

7 BC. The comparison is reported in the supporting information (Fig.S3-S4).

#### *Confined*

Things change dramatically when the matrix representing the confined case ( $C_1$ : cluster 1, in red in the top panel of Fig. 4) is considered. Fig. 5b), d) display the corresponding clustering, in the top panel using the greedy modularity method and in the bottom one using CVS. In this case, the optimal clustering produces 23 spatial clusters with the former approach and 30 with the latter. Both of them predict more clusters, as expected when mobility is reduced. By analyzing the most representative matrices as directed graphs, one can also see that the one for the confined case presents fewer links than the one for non-confined mobility, and that some provinces become singletons in the optimal spatial clustering, see supporting information (Fig.S9 and Fig.S10).

Also in this case both clustering methods provide comparable results: most of the north of Italy is partitioned

similarly with both methods; the singletons (off-white) are essentially the same; also Trento (TN) and Bolzano (BZ), Sassari (SS) and Nuoro (NU), as well as Pescara (PE) and Chieti (CH) are clustered together by pairs with both methods.

## II. CONCLUSIONS

Picking the period 2020-2022 in Italy as a test-case, we introduced a method to assess the main patterns and the most representative movements in a state by using anonymized data from social networks, and showed how this information can be used to identify temporal patterns and spatial communities. The temporal links of the network were inferred from the movement dataset of Facebook users provided by META through its *Data for good* program, combined with publicly available databases from Istat and ISS.

We showed how movement data from social networks can be validated by considering the associated average transition matrix between provinces as the generator of

a Markov jump process, and comparing the corresponding stationary density vector with the population density vector obtained from the official census.

By analyzing the transition matrices time series, we then showed how the normalized probabilities of people moving out of or into provinces collapse on the same average curves, irrespective of the resident populations. Deviations from these two curves can be used to identify particular fluxes due e.g. to tourism.

By considering the distance between transition matrices, we were then able to perform a temporal clustering to distinguish the lockdown periods from the rest. This successfully identifies the first two national lockdowns, which were strictly enforced by the Italian Government. Finally, we picked the most representative transition matrices from the confined (lockdown) and non-confined periods, and used two different methods to identify spatial communities: greedy modularity communities and critical variable selection. The first one finds the communities whose populations move more within a cluster than between clusters; the latter defines clusters to reduce the distance traveled by individuals based on an information-theoretical approach. Both methods return an optimal scale at which actions on circulation in a country could be enforced, and their results are consistent with one another.

As our methodology is completely general, these strategies can be applied to other countries or other scales, as well as different problems relying on a similar kind of data, such as optimizing a transportation network in a city [46], or the analysis of the interaction network between residues to identify coherent or persistent structure in protein dynamics [47].

Finally, we point out that our strategy could also be used to compare the current socio-economical communities in a country to historical data. This can be particularly interesting in regions with a long record of historical documents and whose borders changed significantly over time, of which Italy is a prime example. In this respect, it is noteworthy to observe that the non-confined communities found by modularity bear a strong resemblance with those reported by Pliny the Elder in its *Naturalis Historia* [44], suggesting that the current social, economical, and mobility pattern of Italian communities still echoes its roots dating back by almost two millennia.

### III. MATERIALS AND METHODS

#### A. Datasets

##### *Facebook movement data*

The Facebook (FB) movement data were taken from META's *Data for good* program. The database records the number of people going from province  $i$  to province  $j$ , updated every 8 hour, for Italian users who allowed FB to share such information with the app on their

device; the time frame covered goes from March 1st, 2020 to May 22nd, 2022 (811 days). The database has been completely anonymized by META [48]. In particular, all links between two provinces containing less than 10 people are ignored.

The FB movement data are available both on a grid with cells of roughly  $600 \times 600$  meters at the equator, which is the minimum tile size allowed for privacy protections (Bing tile level 16 [49]), and at the scale of Italian provinces, administrative entities in between municipalities and regions. In this study we concentrate on the province level: the list of 106 provinces used was the official one in 2016 except for the provinces of Sud Sardinia (SU) and Cagliari (CA) which were merged into one node (CA), in order to get inter-compatibility of administrative regions between datasets from FB, Istat, and ISS. A map (Fig.S1) and a table of these provinces can be found in the section 1. of supporting information. In section 2. of section of supporting information describes in detail the workflow of the data preparation.

In this database the FB data reports for each 8 hour period (labeled by  $h$ ):

- The number of FB users moving from province  $i$  to province  $j$  at time  $h$ ,  $n_{ij}^h$  (called  $n_{crisis}$  in the original dataset).
- The total number of FB users in province  $i$  at time  $h$ ,  $n_i^h$ .

##### *Istat and ISS data*

The FB data cover only a fraction of the Italian population (namely those individuals who employ the FB app on mobile devices and enabled location sharing) and does not provide direct information on the population of each province, the amount of COVID cases registered there, or the duration of confinement periods. The population of each province  $i$ ,  $n_i^{\text{Istat}}$ , was obtained from Istat [40], the Italian National Institute of Statistics. We used the most recent database available before the pandemic, released on January 1st, 2020. For simplicity, we assumed that the population remained constant during the period of study: this is an acceptable approximation, given that the global growth rate of the Italian population for that period is roughly  $-0.4\%$  [50] and this fluctuation is negligible for our analysis.

The amount of new COVID cases between February 1st, 2020, and October 7, 2022, reported in the bottom panel of Fig. 4, was obtained from ISS [51]. Data were accumulated as a rolling average over one week.

The dates of the national confinements implemented by the Italian government are the following [25, 52, 53]:

1. from 10/03/2020 to 16/05/2020;
2. from 21/12/2020 to 06/01/2021;

3. from 15/03/2021 to 05/04/2021.

The three periods are indicated by the grey-shaded areas in Figs. 4 and 3. The confinement and de-confinement were progressive processes e.g. at first not all provinces were confined: only two days after the initial, local lockdown the measure was applied to the whole Country. Hence, we chose the temporal boundary of the lockdowns such that the periods correspond to the situation where the whole Country was confined, particularly periods in which any movement between provinces was prohibited.

At smaller scales, national confinements were characterized by rigid restrictions on mobility: in particular, the Italian government provided sanctions of up to three months in prison for those who violated the lockdown, and all non-essential facilities and shops were closed, gyms, swimming pools, spas, wellness centers, museums, cultural centers, ski resorts, cinemas, theatres, pubs, dance schools, game rooms, betting rooms, and bingo halls, discos and similar places in the entire country were suspended. All organized events were also suspended, as well as events in public or private places, counting those of cultural, recreational, sporting, civil, and religious ceremonies, including funeral ceremonies [25].

## B. Stochastic transition matrices

Using the data described in Sec. III A, we built the transition matrices between provinces. As described below, these are averaged daily and over the whole period.

*Mean transition matrix over the whole period*

FB data allowed us to define a mean transition matrix  $\bar{\Pi}$  between nodes as follows:

$$\bar{\Pi}_{ij} = \frac{\sum_h n_{ij}^h}{\sum_j \sum_h n_{ij}^h} \quad (6)$$

where  $\sum_h$  is the sum over all 8-hour-slots during the whole data period. The denominator in Eq.6 normalizes the matrix such that the elements in each row sum to one:  $\sum_j \bar{\Pi}_{ij} = 1, \forall i$ , thus ensuring that  $\bar{\Pi}$  is a stochastic matrix.

*Daily transition matrix*

FB data were used to generate a daily transition matrix representing the link between provinces for each day, indexed by  $t$ . The time evolution of the mobility network was monitored by constructing a time series of transition matrices as follows:

$$\Pi_{ij}(t) = \frac{\sum_{h \in [t-\epsilon, t+\delta]} n_{ij}^h}{\sum_j \sum_{h \in [t-\epsilon, t+\delta]} n_{ij}^h} \quad (7)$$

where  $\sum_{h \in [t-\epsilon, t+\delta]}$  is the sum over all 8-hours-slots in  $[t-\epsilon, t+\delta]$ .

Using Eq. 7 we constructed two different daily time series, one averaged every 24 hours,  $\epsilon = 0$ , and  $\delta = 24$ h, and one based on a weekly rolling average,  $\epsilon = 72$ h days,  $\delta = 96$ h (in between 3 days before and 3 days after day  $t$ ). The weekly averaged one correspond to the average of data provided by ISS.

## C. Temporal Clustering method

To perform the temporal clustering of the transition matrices  $\Pi(t)$ , we used the standard Frobenius matrix norm:

$$\|\Pi(t)\| = \text{Tr}(\Pi(t) \cdot \Pi(t)^T) = \sqrt{\sum_{i=1}^N \sum_{j=1}^N |\Pi(t)_{ij}|^2}, \quad (8)$$

where  $N$  is the number of rows and columns in the transition matrices.

Using this norm we constructed a distance matrix  $D$ , whose elements are the distances between the matrices of the series  $\Pi(t_0), \Pi(t_1), \dots, \Pi(t_T)$ . In other words, for any  $(i, j) \in \llbracket 0, T \rrbracket^2$  an element of  $D$  reads:

$$D_{ij} = \|\Pi(t_i) - \Pi(t_j)\|. \quad (9)$$

We then proceeded to perform an agglomerating clustering with a ward linkage method using the function `sklearn.cluster.AgglomerativeClustering`, available in the `sklearn` Python library [54]. In this bottom-up algorithm, pairs of nodes and then pairs of clusters are recursively merged such that the variance of the distances within the clusters have, for each step, the least possible increase.

The clustering process can be represented in a tree (dendrogram) in which the child branches at each step represent the pairs of clusters that merge into a parent branch. We report this hierarchical clustering dendrogram in Fig. 6. The length of the branches ( $y$ -axis) corresponds to the cophenetic distance, a distance which measures the level of similarity between two merged clusters [55].

The top panel of 6 displays the full dendrogram from individual nodes to one unique cluster. On the bottom panel, this dendrogram is cut at the level of 5 clusters, after which the cophenetic distance increases significantly; the numbers in parentheses (in the  $x$ -axis) are the number of nodes belonging to each cluster.

## D. Spatial Clustering

Spatial clustering into communities are obtained starting from the most representative matrices of the two main



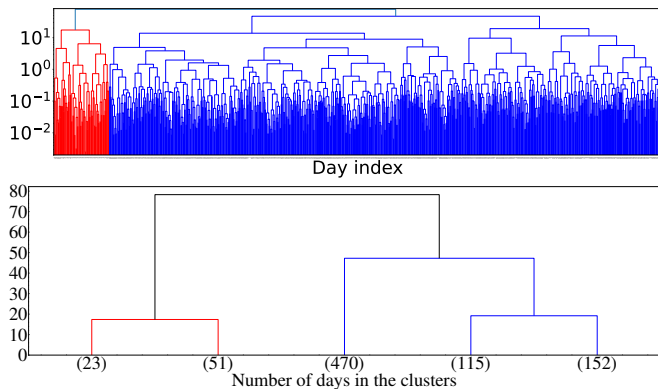


FIG. 6. Hierarchical clustering dendrogram of the day-by-day transition matrices

temporal clusters  $C_0$  and  $C_1$ ; these correspond to the unconfined and confined periods respectively, and are represented in Fig. 4.

#### Most representative current matrices

We computed the mean matrices  $\bar{\Pi}^{C_0}$  and  $\bar{\Pi}^{C_1}$  and the matrices belonging to the unconfined ( $C_0$ ) and confined ( $C_1$ ) temporal clusters. From the mean transition matrices, we selected the most representative ones of each cluster by taking the daily (weekly rolled-average) transition matrix closest to the mean:

$$\tilde{\Pi}^{C_k} = \min_{t \in C_k} \|\Pi(t) - \bar{\Pi}^{C_k}\|, \quad k \in \{0, 1\}, \quad (10)$$

where  $C_k$  is the set of days  $t_i$  within the temporal cluster  $k$ .

The transition matrices defined above provide the daily probability of going from one province to another, but the weights do not contain any information on the population of each province. Hence, using the most representative transition matrices of the two principal temporal clusters and the Istat vector  $\rho^{\text{Istat}}$ , we defined the most representative current matrix  $J^{C_k}$  as follow:

$$J_{ij}^{C_k} = \tilde{\Pi}_{ij}^{C_k} \rho_i^{\text{Istat}}, \quad k \in 0, 1, \quad (11)$$

subject to the normalization condition:

$$\sum_{i,j} J_{ij}^{C_k} = 1. \quad (12)$$

We specify here that we do not define the current matrix using the stationary (Perron-Frobenius) population vector  $\rho^*$  but with the one computed from Istat data which is comparable up to a few fluctuation, as can be seen in Fig. 2. While this means that the detailed balance is not exactly verified, the detailed balance condition is not used in the clustering and the population data of Istat is more accurate, thus ensuring that the computed currents are more representative of the real fluxes.

#### Greedy Modularity Communities method (GMC)

The greedy modularity communities algorithm is provided by the `networkx` Python library (`greedy_modularity_communities`). This algorithm, developed in [56] and refined in [57, 58], relies on the optimization of the modularity  $Q$ . Let  $W_{ij}$  be a weighted matrix, without self-loops, of the associated graph; for a given clustering  $c$ , the modularity is defined as [58]:

$$Q = \frac{1}{2m} \sum_{ij} W_{ij} - \frac{k_i k_j}{2m} \delta(c_i, c_j) \quad (13)$$

where  $m = \frac{1}{2} \sum_{i,j} W_{ij}$  generalises what would be the number of edges in a binary graph,  $k_i = \sum_j W_{ij}$  is the generalised degree of the node  $i$ , and  $c_i$  labels the cluster to which node  $i$  belongs.

To understand its meaning, consider the simpler case of an unweighted graph, where  $W_{ij} = A_{ij}$  is the adjacency matrix. If connections are made at random but respecting the degrees  $k_i$  and  $k_j$  of the nodes  $i$  and  $j$ , then the probability of an existing link between these two nodes is  $k_i k_j / 2m$ . This means that the modularity measures the difference between the linkage of the node within a community cluster and what is expected from a random network. With increasing values of  $Q$ , one has an increasing deviation from a random choice of linkage. Also, looking at Eq.13, we see that if there is only one cluster, then  $\delta(c_i, c_j) \equiv 1$ , and it is straightforward to see that in this case  $Q = 0$ . In the opposite situation, where the clustering is made only of singleton then  $\delta(c_i, c_j) = \delta_{ij}$ ; in this case as well, we see that  $Q = 0$ . It is possible to show [59] that, in between these extreme cases, there exists an optimal clustering corresponding to maximal modularity. The algorithm tests different levels of resolution through an agglomerative clustering method similar to the one presented in section III C, aiming at finding the clustering of the network with maximal modularity.

#### Effective distance matrix between nodes

Following ref [60], we define the effective distance between two adjacent nodes  $i$  and  $j$  as:

$$d_{ij} = 1 - \ln \Pi_{ij}. \quad (14)$$

If there exists a path going from  $i$  to  $j$  with  $l$  steps

$$\Gamma_{ij} = \{(k_0 = i, k_1), (k_1, k_2), \dots, (k_{l-1}, k_l = j)\},$$

the direct length of a path is the sum of the effective distances along its steps:

$$\lambda(\Gamma_{ij}) = \sum_{n=0}^{l-1} d_{k_n, k_{n+1}}. \quad (15)$$

We defined the effective distance as the minimal distance among all the existing paths from  $i$  to  $j$ :

$$D_{ij} = \min_{\Gamma_{ij}} \lambda(\Gamma_{ij}) \quad (16)$$

Then for any two nodes  $i, j$  of the network defined by  $\Pi$ , the effective distance matrix is the symmetric part,

$$\Delta^S = (\Delta + \Delta^t)/2, \quad (17)$$

of the matrix  $\Delta$ , whose elements are defined as follows:

$$\Delta_{ij} = \begin{cases} 0 & \text{if } i = j \\ d_{ij} & \text{if } \Pi_{ij} \neq 0 \\ d_{ji} & \text{if } \Pi_{ij} = 0 \text{ and } \Pi_{ji} \neq 0 \\ D_{ij} & \text{if } \exists \Gamma_{ij} \\ D_{ji} & \text{if } \nexists \Gamma_{ij} \text{ and } \exists \Gamma_{ji} \\ +\infty & \text{elsewhere.} \end{cases} \quad (18)$$

This definition is valid for any weighted directed graph. In particular, the last line is not needed if the graph is weakly connected ( $\forall(i, j), \exists \Gamma_{ij}$  or  $\exists \Gamma_{ji}$ ). Similarly, the two last lines are not needed if it is strongly connected ( $\forall(i, j), \exists \Gamma_{ij}$ ).

In our case, the most representative transition matrix of the non-confined period,  $\bar{\Pi}^{C_0}$  is strongly connected while  $\bar{\Pi}^{C_1}$ , the graph associated with the most representative transition matrix for the confinement period is not even weakly connected, and its connected components are not always strongly connected.

We add that, on a computer, ‘infinite’ must be represented as a large number; this value was defined as 100 times the maximum of the well-defined elements of  $\Delta$ . The effective distance matrix was normalized by its mean value:  $\Delta^S \leftarrow \Delta^S / \bar{\Delta}^S$  where  $\bar{\Delta}^S = \frac{1}{N^2} \sum_{i,j} \Delta_{ij}^S$ . In this way, the agglomerative clustering operations on the distance matrix do not depend on the large-scale cutoff.

#### Critical Variable Selection method (CVS)

The resolution-relevance method [61–66] has been successful in identifying optimal clustering for the reduction of complexity in the representation of biomolecules [67] or for a protein conformational landscape [68].

Considering a set of  $N$  objects and a given clustering of them, we labeled the  $K$  clusters by  $s \in \llbracket 1, K \rrbracket$  and defined  $k_s$  to be the number of objects in cluster  $s$ .  $k_s/N$  is the empirical probability for an object to belong to cluster  $s$ . The *resolution* is defined as the Shannon entropy of this probability distribution:

$$H[s] = - \sum_{s=1}^K \frac{k_s}{N} \log_N \frac{k_s}{N} \quad (19)$$

where  $\log_N$  is the logarithm in base  $N$  such that  $\log_N N = 1$ .  $H[s] = 0$  when all objects belong to only one cluster, and  $H[s] = 1$  at the other extreme, when each object has its own separate cluster.

Resolution alone, however, is not sufficient to identify an optimal level of informativeness of a given clustering. A second quantity, the *relevance*  $H[k]$ , is defined based on the number of clusters containing  $k$  objects,  $m_k$  [43]:

$$m_k = \sum_{s=1}^K \delta_{k, k_s}. \quad (20)$$

The *relevance* is defined as follows:

$$H[k] = - \sum_{k=1}^N \frac{k m_k}{N} \log_N \frac{k m_k}{N}. \quad (21)$$

In the latter expression, the factor  $\frac{k m_k}{N}$  is the empirical probability that a randomly chosen object in the collection belongs to the cluster with  $k$  elements in it. The relevance is the Shannon entropy associated with this second empirical probability.

For both limit cases of 1 and  $N$  clusters,  $H[k] = 0$ , the relevance being non-negative otherwise [43, 68]. The maximum relevance thus corresponds to an optimal clustering, i.e. to the most informative partition of the collection of objects.

We performed an agglomerative clustering of the nodes representing provinces using the distance introduced above, and computed for each number of clusters from 1 to  $N$  the corresponding values of resolution and relevance (see Fig. 7). The optimal partition of provinces was defined as the clustering with the maximum relevance value.

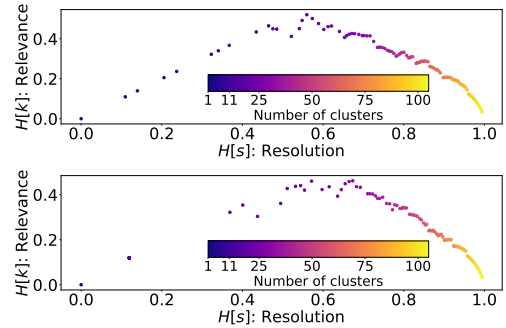


FIG. 7. Resolution versus relevance for agglomerative spatial clustering of temporal cluster  $C_0$  (confined, top) and  $C_1$  (unconfined, bottom).

[1] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whitaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, A. Dighe, J. T. Griffin,

M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley,

- S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson, *Lancet Infect. Dis.* **20**, 669 (2020).
- [2] J. F. Robinson, I. Rios De Anda, and F. J. Moore, *Phys. Fluids* **33**, 43112 (2021).
- [3] S. Talic, S. Shah, H. Wild, D. Gasevic, A. Maharaj, Z. Ademi, X. Li, W. Xu, I. Mesa-Eguiagaray, J. Rostrom, E. Theodoratou, X. Zhang, A. Motee, D. Liew, and D. Ilic, *BMJ* **375**, 10.1136/BMJ-2021-068302 (2021).
- [4] C.-E. Juneau, A.-S. Briand, T. Pueyo, P. Collazzo, and L. Potvin, medRxiv, 2020.07.23.20160234 (2020).
- [5] N. Ahmed, R. A. Michelin, W. Xue, S. Ruj, R. Malaney, S. S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, and S. K. Jha, *IEEE Access* **8**, 134577 (2020), arXiv:2006.10306.
- [6] C. Liu and R. Graham, *Big Data Soc.*, 1 (2021).
- [7] V. Colizza, E. Grill, R. Mikolajczyk, C. Cattuto, A. Kucharski, S. Riley, M. Kendall, K. Lythgoe, D. Bonsall, C. Wymant, L. Abeler-Dörner, L. Ferretti, and C. Fraser, *Nat. Med.* 2021 273 **27**, 361 (2021).
- [8] G. Kostka and S. Habich-Sobiegalla, In times of crisis: Public perceptions toward COVID-19 contact tracing apps in China, Germany, and the United States (2022).
- [9] L. Ricci, D. Di Francesco Maesa, A. Favenza, and E. Ferro, *IEEE Access* **9**, 37936 (2021).
- [10] V. Alfano and S. Ercolano, *Appl. Health Econ. Health Policy* **18**, 509 (2020).
- [11] D. I. Papadopoulos, I. Donkov, K. Charitopoulos, and S. Bishara, *The Impact of Lockdown Measures on COVID-19: A Worldwide Comparison* (2020).
- [12] E. Lavezzo, E. Franchin, C. Ciavarella, G. Cuomo-Dannenburg, L. Barzon, C. Del Vecchio, L. Rossi, R. Manganelli, A. Leregian, N. Navarin, D. Abate, M. Sciro, S. Merigliano, E. De Canale, M. C. Vanuzzo, V. Besutti, F. Saluzzo, F. Onelia, M. Pacenti, S. G. Parisi, G. Carretta, D. Donato, L. Flor, S. Cocchio, G. Masi, A. Sperduti, L. Cattarino, R. Salvador, M. Nicoletti, F. Caldart, G. Castelli, E. Nieddu, B. Labella, L. Fava, M. Drigo, K. A. Gaythorpe, K. E. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, O. Boyd, L. Cattarino, C. Ciavarella, H. L. Coupland, Z. Cucunubá, G. Cuomo-Dannenburg, B. A. Djafaara, C. A. Donnelly, I. Dorigatti, S. L. van Elsland, R. FitzJohn, S. Flaxman, K. A. Gaythorpe, W. D. Green, T. Hallett, A. Hamlet, D. Haw, N. Imai, B. Jeffrey, E. Knock, D. J. Laydon, T. Mellan, S. Mishra, G. Nedjati-Gilani, P. Nouvellet, L. C. Okell, K. V. Parag, S. Riley, H. A. Thompson, H. J. T. Unwin, R. Verity, M. A. Vollmer, P. G. Walker, C. E. Walters, H. Wang, Y. Wang, O. J. Watson, C. Whittaker, L. K. Whittles, X. Xi, N. M. Ferguson, A. R. Brazzale, S. Toppo, M. Trevisan, V. Baldo, C. A. Donnelly, N. M. Ferguson, I. Dorigatti, and A. Crisanti, *Nature* **584**, 425 (2020).
- [13] P. Nouvellet, S. Bhatia, A. Cori, K. E. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, H. Coupland, Z. M. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. A. Djafaara, I. Dorigatti, O. D. Eales, S. L. van Elsland, F. F. Nascimento, R. G. FitzJohn, K. A. Gaythorpe, L. Geidelberg, W. D. Green, A. Hamlet, K. Hauck, W. Hinsley, N. Imai, B. Jeffrey, E. Knock, D. J. Laydon, J. A. Lees, T. Mangal, T. A. Mellan, G. Nedjati-Gilani, K. V. Parag, M. Pons-Salort, M. Ragonnet-Cronin, S. Riley, H. J. T. Unwin, R. Verity, M. A. Vollmer, E. Volz, P. G. Walker, C. E. Walters, H. Wang, O. J. Watson, C. Whittaker, L. K. Whittles, X. Xi, N. M. Ferguson, and C. A. Donnelly, *Nat. Commun.* **12**, 1 (2021).
- [14] A. Glielmo, C. Zeni, B. Cheng, G. Csányi, and A. Laio, *PNAS Nexus* **1**, 1 (2022), arXiv:2104.15079v2.
- [15] J. Wallinga and P. Teunis, *Am. J. Epidemiol.* **160**, 509 (2004).
- [16] G. Grasselli, A. Zangrillo, A. Zanella, M. Antonelli, L. Cabrini, A. Castelli, D. Cereda, A. Coluccello, G. Foti, R. Fumagalli, G. Iotti, N. Latronico, L. Lorini, S. Merler, G. Natalini, A. Piatti, M. V. Ranieri, A. M. Scandroglio, E. Storti, M. Cecconi, and A. Pesenti, *JAMA - J. Am. Med. Assoc.* **323**, 1574 (2020).
- [17] E. Bertuzzo, L. Mari, D. Pasetto, S. Miccoli, R. Casagrandi, M. Gatto, and A. Rinaldo, *Nat. Commun.* **11**, 1 (2020).
- [18] M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, and A. Rinaldo, *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10484 (2020).
- [19] V. Marziano, G. Guzzetta, B. M. Rondinone, F. Boccuni, F. Riccardo, A. Bella, P. Poletti, F. Trentini, P. Pezzotti, S. Brusaferrò, G. Rezza, S. Iavicoli, M. Ajellid, and M. Stefano, *Proc. Natl. Acad. Sci. U. S. A.* **118**, 10.1073/PNAS.2019617118/-/DCSUPPLEMENTAL (2021).
- [20] C. Urzeala, M. Duclos, U. Chris Ugboe, A. Bota, M. Berthon, K. Kulik, D. Thivel, R. Bagheri, Y. Gu, J. S. Baker, N. Andant, B. Pereira, K. Rouffiac, M. Clinchamps, F. Duthheil, S. Mestres, C. Miele, V. Navel, L. Parreira, Y. Boirie, J. B. Bouillon-Minois, M. L. Fantini, J. Schmidt, S. Tubert-Jeannin, P. Chausse, M. Dambun, S. Droit-Volet, J. Guegan, S. Guimond, L. Mondillon, A. Nugier, P. Huguët, S. Dewavrin, F. Marhar, G. Naughton, A. Benson, C. Lamm, V. Drapeau, R. Avilés Dorlhiac, B. Bustos, H. Zhang, P. Dieckmann, B. Quach, Y. Duan, G. Gao, W. Y. Huang, K. L. K. Lau, C. Q. Zhang, J. Jiao, K. Chou Chen, H. Nasir, P. Cocco, R. Lecca, M. Puligheddu, M. Figorilli, M. Charkhabi, D. Pfabigan, P. Dieckmann, S. Antunes, D. Neto, P. Almeida, M. J. Gouveia, P. Quinteiro, B. Dubuis, J. Lemaignan, A. Liu, and F. Saadaoui, *Heal. Expect.* **25**, 522 (2022).
- [21] M. R. Gualano, G. Lo Moro, G. Voglino, F. Bert, and R. Siliquini, *Int. J. Environ. Res. Public Health* **17**, 1 (2020).
- [22] M. Natilli, A. Rossi, A. Trecroci, L. Cavaggioni, G. Merati, and D. Formenti, *Sci. Data* **9**, 1 (2022).
- [23] ISS, *Faq sul calcolo del rt*, [https://www.iss.it/en/coronavirus/-/asset\\_publisher/1SRKHcCJJQ7E/content/faq-sul-calcolo-del-rt](https://www.iss.it/en/coronavirus/-/asset_publisher/1SRKHcCJJQ7E/content/faq-sul-calcolo-del-rt) (2020).
- [24] *Gazzetta Ufficiale*, Decreto-legge 23 febbraio 2020, n.6, [https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2020-02-23&atto.codiceRedazionale=20G00020&tipoSerie=serie\\_generale&tipoVigenza=originario](https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2020-02-23&atto.codiceRedazionale=20G00020&tipoSerie=serie_generale&tipoVigenza=originario) (2020).
- [25] *Gazzetta Ufficiale*, Decreto-legge 2 marzo 2020, n.9, [https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2020-03-02&atto.codiceRedazionale=20G00026&tipoSerie=serie\\_generale&tipoVigenza=originario](https://www.gazzettaufficiale.it/atto/vediMenuHTML?atto.dataPubblicazioneGazzetta=2020-03-02&atto.codiceRedazionale=20G00026&tipoSerie=serie_generale&tipoVigenza=originario) (2020).
- [26] E. Leoni, G. Cencetti, G. Santin, T. Istomin, D. Molteni, G. P. Picco, E. Farella, B. Lepri, and A. L. Murphy, *EPJ*

- Data Sci. **11**, 10.1140/epjds/s13688-022-00316-y (2022), arXiv:2106.14750.
- [27] J. Fournet and A. Barrat, PLoS One **9**, 10.1371/journal.pone.0107878 (2014).
- [28] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. F. Pinton, M. Quaggiotto, W. van den Broeck, C. Régis, B. Lina, and P. Vanhems, PLoS One **6**, 10.1371/JOURNAL.PONE.0023176 (2011).
- [29] A. Barrat, C. Cattuto, V. Colizza, F. Gesualdo, L. Isella, E. Pandolfi, J. F. Pinton, L. Ravà, C. Rizzo, M. Romano, J. Stehlé, A. E. Tozzi, and W. Van den Broeck, Eur. Phys. J. Spec. Top. **222**, 1295 (2013).
- [30] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-f. Pinton, N. Khanafer, V. D. Broeck, and P. Vanhems, BMC Med. **9**, 1 (2011).
- [31] D. A. Contreras, E. Colos, G. Bassignana, V. Colizza, and A. Barrat, J. R. Soc. Interface **19**, 10.1098/rsif.2022.0164 (2022).
- [32] V. Colizza, R. Pastor-Satorras, and A. Vespignani, Nat. Phys. **3**, 276 (2007), arXiv:0703129 [cond-mat].
- [33] V. Colizza and A. Vespignani, J. Theor. Biol. **251**, 450 (2008), arXiv:0706.3647.
- [34] H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. Vollmer, C. Whittaker, S. L. Filippi, X. Xi, M. Monod, O. Ratmann, M. Hutchinson, F. Valka, H. Zhu, I. Hawryluk, P. Milton, K. E. Ainslie, M. Baguelin, A. Boonyasiri, N. F. Brazeau, L. Cattarino, Z. Cucunuba, G. Cuomo-Dannenburg, I. Dorigatti, O. D. Eales, J. W. Eaton, S. L. van Elsland, R. G. FitzJohn, K. A. Gaythorpe, W. Green, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, J. Lees, G. Nedjati-Gilani, P. Nouvellet, L. Okell, K. V. Parag, I. Siveroni, H. A. Thompson, P. Walker, C. E. Walters, O. J. Watson, L. K. Whittles, A. C. Ghani, N. M. Ferguson, S. Riley, C. A. Donnelly, S. Bhatt, and S. Flaxman, Nat. Commun. **11**, 1 (2020).
- [35] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, Bull. Math. Biol. **68**, 1893 (2006).
- [36] T. M. Le, L. Raynal, O. Talbot, H. Hambridge, C. Drovandi, A. Mira, K. Mengersen, and J. P. Onnela, Sci. Rep. **12**, 10.1038/s41598-022-10678-y (2022).
- [37] D. Balcan, V. Colizza, B. Gonçalves, H. Hud, J. J. Ramasco, and A. Vespignani, Proc. Natl. Acad. Sci. U. S. A. **106**, 21484 (2009).
- [38] C. Zhong, R. Morphet, and M. Yoshida, PLoS One **18**, e0284902 (2023).
- [39] Meta Company, Data for good program, <https://dataforgood.facebook.com/> (2023).
- [40] Italian National Institute of Statistics, Istat, <https://www.istat.it/en/> (2023).
- [41] ISTAT, Previsioni della popolazione residente base 1.1.2021 nota metodologica, [https://demo.istat.it/data/previsioni/nota\\_previsioni\\_demografiche\\_demo.pdf](https://demo.istat.it/data/previsioni/nota_previsioni_demografiche_demo.pdf) (2021).
- [42] T. P. Smith, S. Flaxman, A. S. Gallinat, S. P. Kinoshian, M. Stemkovski, H. Juliette, O. J. Watson, C. Whittaker, L. Cattarino, I. Dorigatti, M. Tristem, and W. D. Pearse, Proc. Natl. Acad. Sci. U. S. A. **118**, e2019284118 (2021).
- [43] M. Marsili, I. Mastromatteo, and Y. Roudi, J. Stat. Mech. Theory Exp. **2013**, 9003 (2013).
- [44] Pliny the Elder, Naturalis historia iii. 46, [http://www.attalus.org/translate/pliny\\_hn3a.html](http://www.attalus.org/translate/pliny_hn3a.html) (77 CE).
- [45] G. Cifani, Soc. Evol. Hist. **9**, 53 (2010).
- [46] S. Bontorin, G. Cencetti, R. Gallotti, B. Lepri, and M. De Domenico, Prepr. arXiv (2023), arXiv:2301.08661.
- [47] M. Tiberti, G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber, and E. Papaleo, J. Chem. Inf. Model. **54**, 1537 (2014).
- [48] Meta Company, Protecting privacy in facebook mobility data during the covid-19 response, <https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response> (2020).
- [49] Microsoft Company, Bing maps tile system, <https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system> (2022).
- [50] MacroTrends, Italy population growth rate 1950-2023, <https://www.macrotrends.net/countries/ITA/italy/population-growth-rate> (2023).
- [51] Italian National Institute for Health, Iss, <https://www.iss.it/web/iss-en> (2023).
- [52] Gazzetta Ufficiale, Decreto-legge 2 dicembre 2020, n.158, <https://www.gazzettaufficiale.it/eli/id/2020/12/02/20G00184/sg> (2020).
- [53] Gazzetta Ufficiale, Decreto-legge 13 marzo 2021, n.30, <https://www.gazzettaufficiale.it/eli/id/2021/03/13/21G00040/sg> (2021).
- [54] Scikit learn Python Library, sklearn.cluster.agglomerativeclustering, <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html> (2022).
- [55] SciPy Python Library, scipy.cluster.hierarchy.cophenet, <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.cluster.hierarchy.cophenet.html> (2022).
- [56] M. E. Newman, Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. **69**, 5 (2004).
- [57] A. Clauset, M. E. Newman, and C. Moore, Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. **70**, 6 (2004), arXiv:0408187 [cond-mat].
- [58] M. E. Newman, Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. **70**, 9 (2004).
- [59] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hofer, Z. Nikoloski, and D. Wagner, IEEE Trans. Knowl. Data Eng. **20**, 172 (2008).
- [60] D. Brockmann and D. Helbing, Science (80-. ). **342**, 1337 (2013).
- [61] O. Obregón, J. L. López, and M. Ortega-Cruz, Entropy **20**, 755 (2018).
- [62] R. Cubero, M. Marsili, and Y. Roudi, Entropy **20**, 755 (2018).
- [63] R. J. Cubero, J. Jo, M. Marsili, Y. Roudi, and J. Song, J. Stat. Mech. Theory Exp. **2019**, 063402 (2019), arXiv:1808.00249.
- [64] R. J. Cubero, M. Marsili, and Y. Roudi, J. Comput. Neurosci. **48**, 85 (2020), 1802.10354.
- [65] M. Marsili and Y. Roudi, Quantifying relevance in learning and inference (2022), 2202.00339.
- [66] R. Holtzman, M. Giulini, and R. Potestio, Phys. Rev. E **106**, 10.1103/PhysRevE.106.044101 (2022).
- [67] M. Giulini, R. Menichetti, M. S. Shell, and R. Potestio, J. Chem. Theory Comput. **16**, 6795 (2020), 2004.03988.
- [68] M. Mele, R. Covino, and R. Potestio, Soft Matter **18**, 7064 (2022).

## 1. Supplementary Information

## 2. List of Provinces used in the study

node index	Provinces name	Car plate code	ITTER107 code	Population 01 Jan 2020	node index	Provinces name	Car plate code	ITTER107 code	Population 01 Jan 2020
0	Agrigento	AG	ITG14	412427.0	53	Mantova	MN	ITC4B	404440.0
1	Alessandria	AL	ITC18	407049.0	54	Modena	MO	ITD54	702787.0
2	Ancona	AN	ITE32	461745.0	55	Massa-Carrara	MS	ITE11	188395.0
3	Aosta	AO	ITC20	123337.0	56	Matera	MT	ITF52	191663.0
4	Ascoli Piceno	AP	ITE34	202317.0	57	Napoli	NA	ITF33	2967117.0
5	L'Aquila	AQ	ITF11	288439.0	58	Novara	NO	ITC15	361845.0
6	Arezzo	AR	ITE18	334634.0	59	Nuoro—Ogliastra	NU	ITG26	199349.0
7	Asti	AT	ITC17	207939.0	60	Oristano	OR	ITG28	150812.0
8	Avellino	AV	ITF34	399623.0	61	Palermo	PA	ITG12	1199626.0
9	Bari	BA	ITF42	1224756.0	62	Piacenza	PC	ITD51	283889.0
10	Bergamo	BG	ITC46	1102670.0	63	Padova	PD	ITD36	930898.0
11	Biella	BI	ITC13	169560.0	64	Pescara	PE	ITF13	313346.0
12	Belluno	BL	ITD33	198518.0	65	Perugia	PG	ITE21	641318.0
13	Benevento	BN	ITF32	263460.0	66	Pisa	PI	ITE17	417245.0
14	Bologna	BO	ITD55	1015701.0	67	Pordenone	PN	ITD41	310158.0
15	Brindisi	BR	ITF44	379851.0	68	Prato	PO	ITE15	264397.0
16	Brescia	BS	ITC47	1254322.0	69	Parma	PR	ITD52	450044.0
17	Barletta-Andria-Trani	BT	IT110	379251.0	70	Pistoia	PT	ITE13	289256.0
18	Bolzano	BZ	ITD10	535774.0	71	Pesaro e Urbino	PU	ITE31	351993.0
19	Cagliari—Sud Sardegna	CA	ITG27—IT111	754878.0	72	Pavia	PV	ITC48	534691.0
20	Campobasso	CB	ITF22	210599.0	73	Potenza	PZ	ITF51	348336.0
21	Caserta	CE	ITF31	900293.0	74	Ravenna	RA	ITD57	386007.0
22	Chieti	CH	ITF14	372473.0	75	Reggio di Calabria	RC	ITF65	518978.0
23	Caltanissetta	CL	ITG15	250550.0	76	Reggio nell'Emilia	RE	ITD53	524193.0
24	Cuneo	CN	ITC16	580789.0	77	Ragusa	RG	ITG18	315082.0
25	Como	CO	ITC42	594657.0	78	Rieti	RI	ITE42	150689.0
26	Cremona	CR	ITC4A	351287.0	79	Roma	RM	ITE43	4222631.0
27	Cosenza	CS	ITF61	671171.0	80	Rimini	RN	ITD59	336916.0
28	Catania	CT	ITG17	1068835.0	81	Rovigo	RO	ITD37	229097.0
29	Catanzaro	CZ	ITF63	341991.0	82	Salerno	SA	ITF35	1060188.0
30	Enna	EN	ITG16	155982.0	83	Siena	SI	ITE19	262046.0
31	Forlì-Cesena	FC	ITD58	391524.0	84	Sondrio	SO	ITC44	178208.0
32	Ferrara	FE	ITD56	340755.0	85	La Spezia	SP	ITC34	214879.0
33	Foggia	FG	ITF41	597902.0	86	Siracusa	SR	ITG19	383743.0
34	Firenze	FI	ITE14	994717.0	87	Sassari—Olbia-Tempio	SS	ITG25	474142.0
35	Fermo	FM	IT109	168485.0	88	Savona	SV	ITC32	267748.0
36	Frosinone	FR	ITE45	468438.0	89	Taranto	TA	ITF43	558130.0
37	Genova	GE	ITC33	816250.0	90	Teramo	TE	ITF12	299402.0
38	Gorizia	GO	ITD43	138666.0	91	Trento	TN	ITD20	542158.0
39	Grosseto	GR	ITE1A	216989.0	92	Torino	TO	ITC11	2205104.0
40	Imperia	IM	ITC31	208561.0	93	Trapani	TP	ITG11	415233.0
41	Isernia	IS	ITF21	80170.0	94	Terni	TR	ITE22	218254.0
42	Crotone	KR	ITF62	161744.0	95	Trieste	TS	ITD44	230623.0
43	Lecco	LC	ITC43	332435.0	96	Treviso	TV	ITD34	876755.0
44	Lecce	LE	ITF45	772276.0	97	Udine	UD	ITD42	517848.0
45	Livorno	LI	ITE16	326716.0	98	Varese	VA	ITC41	878059.0
46	Lodi	LO	ITC49	227064.0	99	Verbano-Cusio-Ossola	VB	ITC14	154233.0
47	Latina	LT	ITE44	565840.0	100	Vercelli	VC	ITC12	165760.0
48	Lucca	LU	ITE12	381890.0	101	Venezia	VE	ITD35	839396.0
49	Monza e Brianza	MB	IT108	870112.0	102	Vicenza	VI	ITD32	852861.0
50	Macerata	MC	ITE33	305249.0	103	Verona	VR	ITD31	927108.0
51	Messina	ME	ITG13	599990.0	104	Viterbo	VT	ITE41	307592.0
52	Milano	MI	ITC45	3237101.0	105	Vibo Valentia	VV	ITF64	150702.0

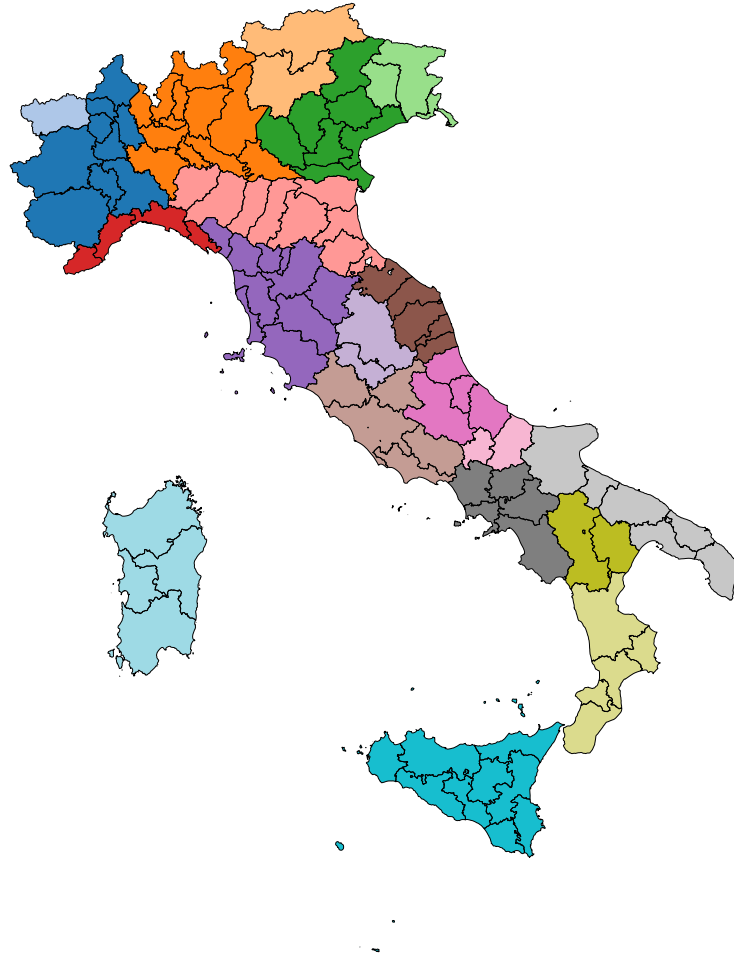


FIG. SI.1. Administrative borders of the provinces of Italy considered in the study (black lines) and of the regions (in colors) as there are defined now.

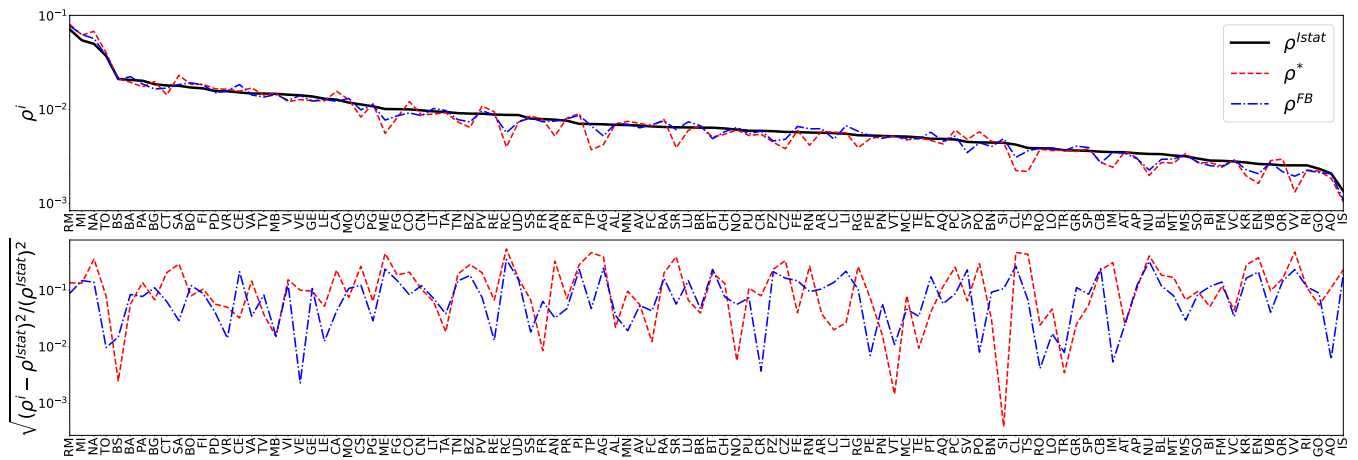


FIG. SI.2. Top Panel: Comparison of the different population density vector from Facebook and Istat data:  $\rho^{Istat}$ ,  $\rho^{FB}$  and  $\rho^*$  against provinces order by Istat population. Bottom Panel: Standard deviation of the Facebook data vectors from the Istat data vector is displayed showing a good agreement between datasets.

### 3. Region of Italy organized by Emperor Augustus



FIG. SI.3. The proposed partition of Emperor Augustus, together with the optimal spatial clustering for the non-confided periods (left with GMC, right with CVS). Source central map : [https://it.wikipedia.org/wiki/Regioni\\_dell%27Italia\\_augustea#/media/File:Regioni\\_dell'Italia\\_Augustea.svg](https://it.wikipedia.org/wiki/Regioni_dell%27Italia_augustea#/media/File:Regioni_dell'Italia_Augustea.svg)

## 4. Data Preparation

### 5. Data Sources

Mainly two data sources were used for our analysis: the **Facebook Data for Good Italy Coronavirus Disease Prevention Map** data and the **COVID-19 data** published by the ISS (Istituto Superiore di Sanità - National Institute of Health) - INFN (Istituto Nazionale di Fisica Nucleare - National Institute for Nuclear Physics) collaboration.

Respectively, the two data sources contained the following datasets that were of interest to us:

- The **Movements Between Administrative Regions** dataset describes the number of Facebook users that move between two NUTS-3 administrative regions (aka *province*). The temporal aggregation of the dataset is of 8 hours, meaning that if a person is checked-in in region A in a certain time frame, and the same person is found to be checked-in in another region B in the subsequent time frame, then a movement between regions A and B are counted. A 24-hour the day is divided into three time frames: 00:00-08:00, 08:00-16:00 and 16:00-24:00. Two types of data are considered: the baseline, which is computed by taking the average on the same weekday for the same weekdays, and the people during the crisis, which is the actual number of people detected in the specified DateTime. Only users of the Facebook app that have the Location History option enabled are counted, and also if the aggregation yields counts under 10 units then the datum is discarded.
- The **New Positive Cases By Date** dataset describes the number of new positive SARS-CoV-2 cases, aggregated by date and province of detection. This dataset does not suffer from the lag between detection and publication, unlike the Dipartimento della Protezione Civile (Department for Civil Defense) data. The number of cases is the result of a window average over a week, where the final result is the day in the middle of the week (the fourth day of the week).

### 6. The choice of the stack

In order to perform the extraction, loading, and transformation of the data various paths have been explored, but our choice fell on the current technological stack.

- **Python** is the main scripting language and piece of software used throughout the whole pipeline. Its user-friendliness, its widespread use among both the industry and researchers, and the availability of great libraries for data science and visualization made it our natural choice for our purposes. In particular, the libraries mainly used by us are Selenium (web browser automation) and Pandas (data analysis and manipulation).
- **Miller** is a toolkit for data munging. It allows quick CSV manipulations and it contains several powerful commands, that can also be chained one after the other.
- **Bash** is used for integrating and preprocessing various data sources. It is extremely flexible and compatible with most of Unix-like systems, and for certain types of data science workloads it can quickly and efficiently get the work done.
- **DuckDB** is an embeddable analytical database. It is similar to SQLite, in that the database system runs within a host process, but it is optimized for analytical (OLAP) workloads. It allows for manipulations *à la* Pandas but also has full-query optimization and transactional storage. It is a good choice for our purposes since it allows faster queries to be made, it does not require the maintenance of a DB stack and it integrates very well with Python thanks to the DuckDB Python API.

### 7. Gathering the data

The **Facebook Data for Good** data can only be downloaded by using an online interface, and each transaction is size-capped (i.e. Movements Between Administrative Regions data for more than a two-weeks period would not be downloaded). In order to facilitate and speed up the sourcing of the data, a bulk download tool was developed. The tool makes use of Selenium, a Python library for browser automation, that allows automated workflows that simulate human interaction with a browser. The resulting raw data are available as zip archives containing many CSV files. The **COVID-19 data** by ISS/INFN is published as a single zip archive containing many CSV files, one for each aggregation, data type and province/region.



Reference tables are also essential for the analysis, as they allow data integration between incoherent definitions (in particular, concerning spatial aggregation units) between different datasets. Some of them are manually compiled, and others are aggregate data extracted from the original datasets:

- The **provinces** identifications conversion table has been created manually. It contains the correspondence between IDs for provinces in the Facebook dataset, the ISS/INFN dataset, and the “car number plate code” two letters characters).
- The **locations** reference table contains the latitude and longitude for each province.

## 8. Cleaning and loading the data

The raw data is then loaded into our database for further analysis.

First of all, the archives are unpacked and the data is cleaned for our purposes with the use of Miller. The following operations are performed:

- The data is filtered in order to get only data for Italian provinces (generally Facebook uses rectangular bounding boxes to get subsets of data).
- Minor changes in data formats are operated (such as missing date-time imputation and format correction for date-time strings).
- Null entries are discarded.
- Only columns of our interest are selected.

Then the data are piped through an SQL COPY command, that loads it in a DuckDB database.

## 9. Transforming the data

This is the last step of our data preparation pipeline.

In this step, we transform the raw data contained in the database into SQL tables with SQL views, in such a way that they can be accessed easily and are expressed in a manner that is optimal for the analysis purposes of our research.

The Movements Between Administrative Regions dataset has rows aggregated and summed over with the new definitions of provinces as defined in the reference table.

Starting from this table, then multiple views are created:

- total number of people moving from each origin place by date;
- total number of people moving between places summed over by each day;
- daily probability that a movement between places happens (aka *transition matrix*);
- total probability that a movement between places happens;
- weekly rolling average of the daily probability of movement between places.

The COVID-19 ISS row dataset is also aggregated and summed over the province using the new definitions as defined in the reference table. 2.

## 10. About Perron-Frobenius (PF) theorem for stochastic matrix

In the graph identified by the mean matrix  $\bar{\Pi}$  there is a non-zero probability to reach any node from any other node in a finite number of steps, that is, the graph is strongly connected and aperiodic. (To say a graph is aperiodic is equivalent to saying that its representative matrix is irreducible or saying that the random walk on the graph is ergodic.) Then, the transition matrix representing the graph is non-negative and irreducible.

For a general non-negative irreducible matrix,  $\mathbf{\Pi}$ , the PF theorem then ensures that the highest eigenvalue  $\lambda^*$  of  $\mathbf{\Pi}$  is not degenerate. It is often called the PF eigenvalue and we name PF left eigenvector  $l^*$  (and right eigenvector  $r^*$ ) its associated eigenvectors.

A consequence of the theorem in this case is also that any distribution on which we apply the matrix successively, will concentrate, in the long time (long path limit), to the stationary density vector  $\rho_i^* = l_i^* r_i^*$ .

We explain it here in our specific case where the matrix is stochastic.

The normalization of  $\mathbf{\Pi}$  is such that it is stochastic, i.e. each its rows sum to 1, i.e 1 is eigenvalue of  $\bar{\Pi}$  and is associated with the right eigenvector  $\mathbf{r}^* = \mathbf{1}_{np} = (1, 1, \dots, 1)^T$  and  $l^*$  :

$$l^* = l^* \bar{\Pi} \quad (22)$$

Moreover, one can also show that in this case, 1 is the maximum eigenvalue possible and the PF theorem ensure it is unique, as well as its associated eigenvectors.

Therefore, for stochastic matrix, we commonly identify  $\rho^* = l^*$  as the stationary density vector but in general,  $r_i^*$  may be something else than the  $\mathbf{1}$  and the PF eigenvalue different than 1.

In the following, we show the long-time limit convergence in our case.

Because the PF left eigenvector of  $\mathbf{\Pi}$  is the unique invariant, it ensures the detail balance of the associated Markov process:

$$\rho_i^* \Pi_{ij} = \Pi_{ji} \rho_j^*$$

. Multiplying the left and the right by  $\rho_i^{*-1/2}$  and  $\rho_j^{*-1/2}$ , we obtain that

$$\rho_i^{*1/2} \Pi_{ij} \rho_j^{*-1/2} = \rho_i^{*-1/2} \Pi_{ji} \rho_j^{*1/2} = (\rho_j^{*-1/2} \Pi_{ij} \rho_i^{*1/2})^T. \quad (23)$$

Hence the matrix  $S$  defined by the elements

$$S_{ij} = \rho_i^{*1/2} \Pi_{ij} \rho_j^{*-1/2}, \quad (24)$$

is equal to its own transposed, and so is symmetric.

Defining the transformation  $U = \text{diag}(\rho_i^{*1/2})$  we have:

$$S = U^{-1} \mathbf{\Pi} U, \quad (25)$$

then  $S$  have the same eigenvalues (for  $S$ , left and right eigenvector are the same) and is symmetric, so diagonalizable in real space, so  $\mathbf{\Pi}$  itself is diagonalizable.

So there exists a mapping, i.e. a base of the vector space,  $O$  such that :

$$\mathbf{\Pi} = O^{-1} \mathbf{\Lambda} O$$

with  $\mathbf{\Lambda} = \text{diag}(\lambda^*, \lambda_1, \dots, \lambda_N)$ .

The PF theorem ensures that for our stochastic matrix  $\lambda^* = 1 > |\lambda_i| \geq 0$ ,  $i = 1, \dots, N$ , Hence, in the long time limit (or long path limit).

$$\mathbf{\Lambda}^t \sim \text{diag}(1, 0, 0, 0, \dots, 0)$$

The principal (or Perron-Frobenius) eigenvalue  $\lambda^* = 1$  will dominate and any non-trivial distribution  $\rho$  over the nodes will converge to the unique stationary density vector:

$$\rho \mathbf{\Pi}^t \sim \rho^*.$$

## 11. Clustering of the mean current Matrix

In Fig.SI.4, on the top is displayed a representation of the mean current matrix sorted by clusters and by weight, one sees that the method is satisfactory giving well-defined blocks corresponding to each community. On the bottom panel, we see that the clusters correspond, apart from very few border cases,(and Umbria is split apart) to a group of regions in Italy. In detail, for the ten clusters found, we have:

- The green cluster corresponds perfectly to the “Triveneto” region (that is, Veneto, Friuli-Venezia-Giulia, and the provinces of Trento and Bolzano).
- The red one to Lombardia with the exception of Mantova plus the two provinces of Verbano/Cusio/Ossola and Novara (belonging to Piemonte) and Piacenza (belonging to Emilia-Romagna).
- The dark purple corresponds to the region of Valle d’Aosta and Piemonte (minus VB and NO) and Liguria, at the exception of Spezia.
- The yellow cluster is Toscana plus the provinces of Spezia (Liguria) and Perugia (Umbria)
- The orange one corresponds to the region of Emilia-Romagna at the exception of Piacenza (PC) and adds the provinces of Pesaro/Urbino (Marche) and Mantova (Lombardia).
- The grey cluster is the regions of Marche (minus PU), Abruzzo (minus AQ) and Molise (minus IS)
- The teal one matches with the regions of Lazio and Sardegna (plus TE (Umbria) and AQ (Abruzzo))
- The light purple corresponds to the region of Campania plus the province of Isernia belonging to Molise.
- The light blue cluster corresponds perfectly o the regions of Puglia and Basilicata
- Finally, the blue cluster perfectly to the regions of Calabria and Sicilia.

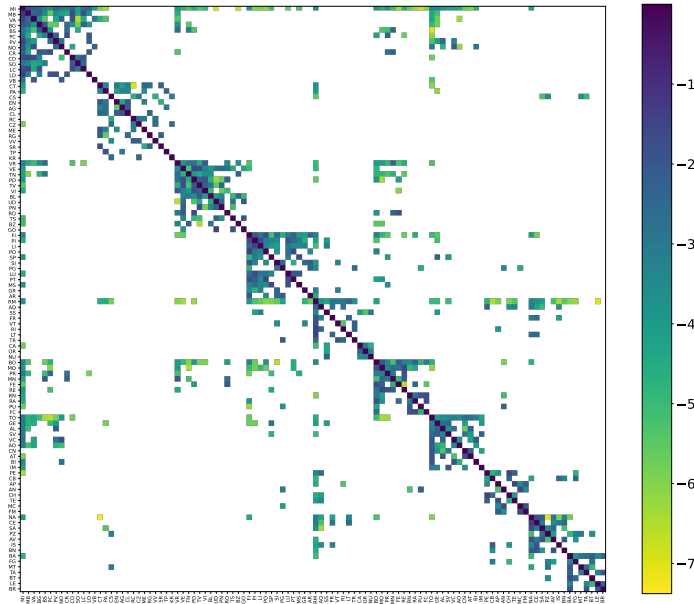
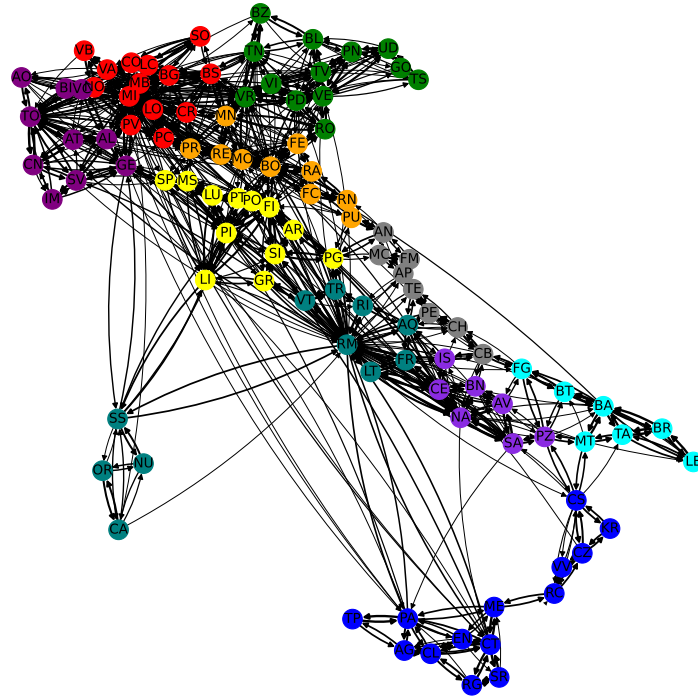


FIG. SI.4. Italian provinces clustered using GMC by communities using the all time-averaged matrix.

Top panel: Representation of the clustered mean current matrix, for visualization the shades are in  $\log_{10}$  of the mean probabilities of going from one province to another.

Bottom: Graph representation of the community clustering with colors corresponding to the different clusters, the widths of the links are proportional to the logarithm of the transition probability. At the exception of Sardinia that is in Rome cluster here, the clustering is the same than non-confined most representative matrix using GMC.

12. Z-score of the probability of going in and out of provinces

To better see which provinces differ the most from the mean trend we compute the Z-score, which is defined for the time series  $X_i(t)$  of province  $i$  as:

$$Z_i = \frac{1}{T} \sum_{t=0}^T \frac{|X_i(t) - \mu(t)|}{\sigma(t)} \tag{26}$$

with  $\mu(t) = \langle X_i(t) \rangle$  being the average over provinces and  $\sigma(t) = \sqrt{\langle X_i(t) - \mu(t) \rangle}$  the standard deviation at time  $t$ .

In Fig. SI.5, we show at the top a map with the Z score for the outgoing probabilities, and on the bottom one for ingoing probabilities for each province.

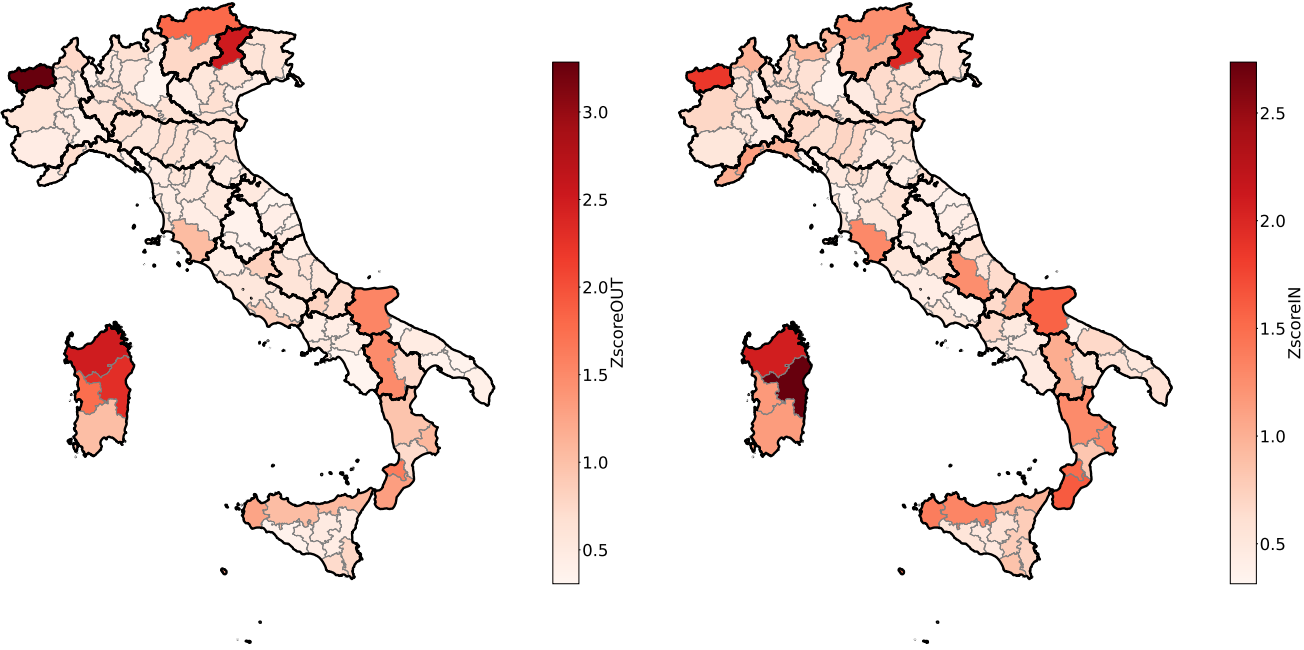


FIG. SI.5. Map of the 2-year average Z score by province. The Z scores defined Eq. 26 are the average fluctuations by provinces of the probability of moving in (Z score IN) and out (Z score OUT) of the province.

### 13. In and outgoing probability for each spatial cluster

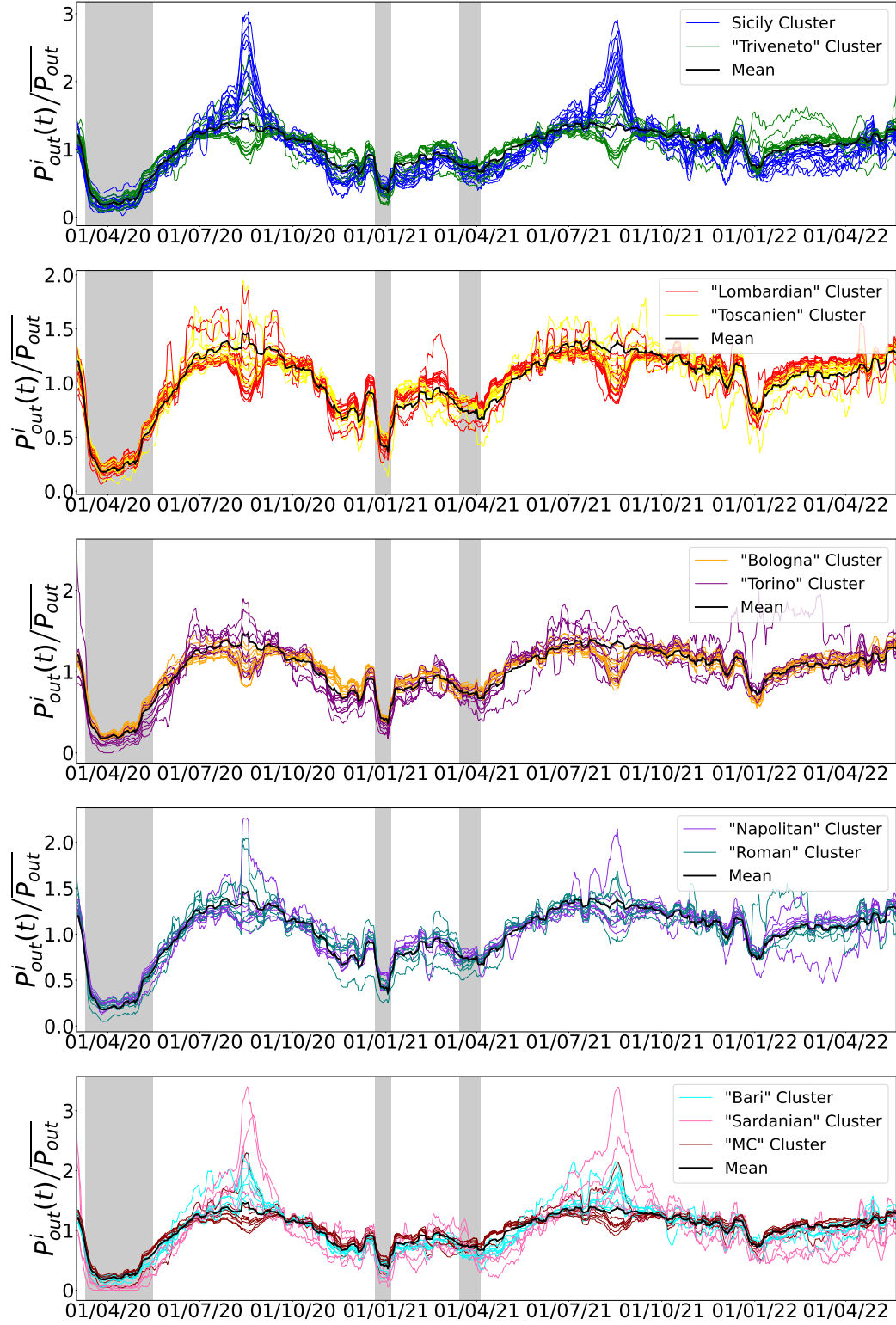


FIG. SI.6. Outgoing probability for each cluster, the colors correspond to the clusters of the non-confined case using GMC.

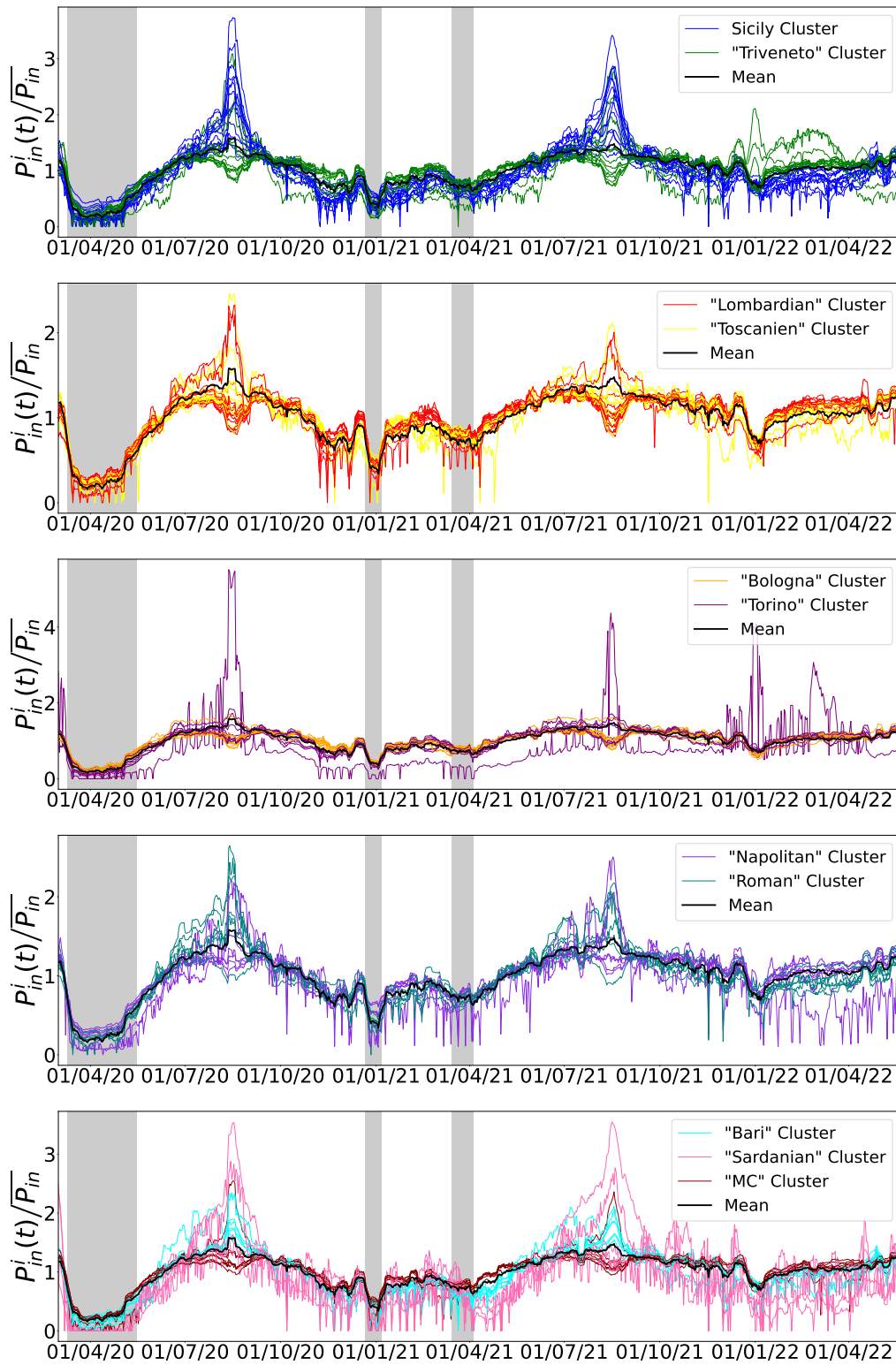


FIG. SI.7. Incoming probability for each cluster the colors correspond to the clusters of the non-confined case using GMC .

14. Network representation of the spatial partition with the two clustering methods

We display here the full network visualization of the two most representatives obtained for the two temporal clusters, the widths of the links are proportional to the logarithm of the transition probability.

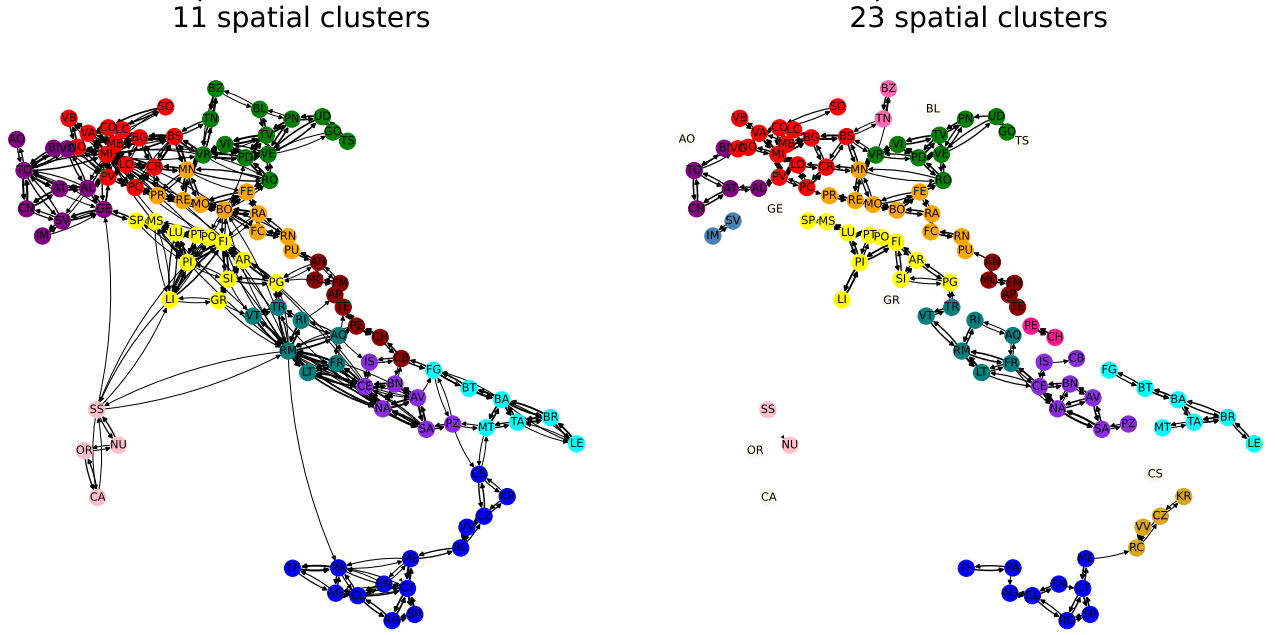


FIG. SI.8. Directed graph representation of the most representative matrices for non-confined cluster  $C_0$  (top) and confined one  $C_1$  (bottom) the optimal clustering in using the greedy modularity method.

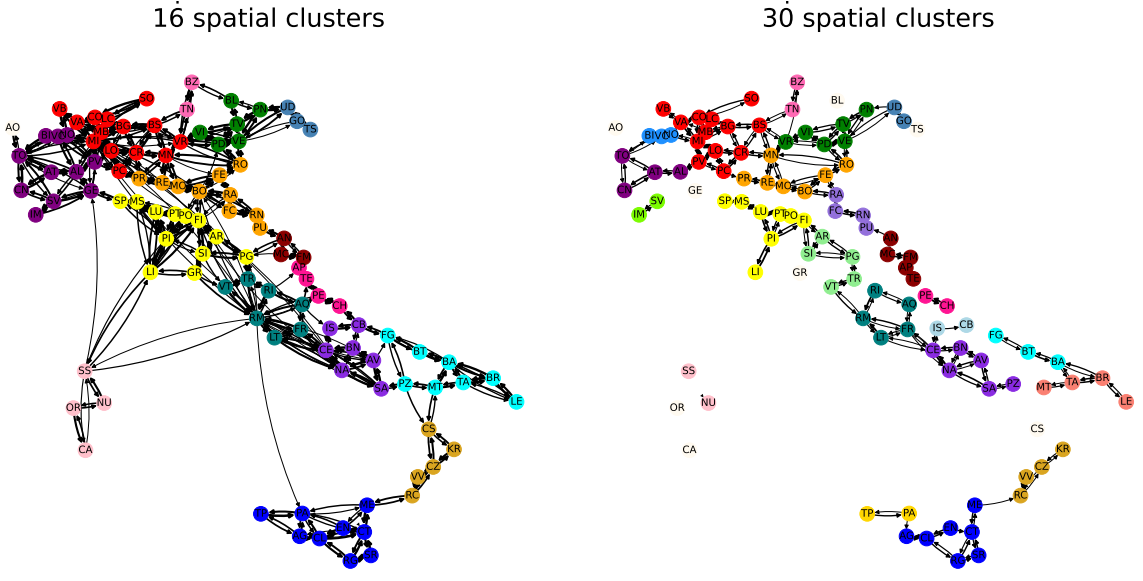


FIG. SI.9. Directed graph representation of the most representative matrices for non-confined cluster  $C_0$  (top) and confined one  $C_1$  (bottom) the optimal clustering in using the critical variable selection method.