Trends in Computational and Applied Mathematics, **23**, N. 4 (2022), 705-729
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2022.023.04.00705

A Spectral Clustering Approach for the Evolution of the COVID-19 Pandemic in the State of Rio Grande do Sul, Brazil

L. E. ALLEM[1*], C. HOPPEN[2], M. M. MARZO[3] and L. S. SIBEMBERG[4]

Received on August 1, 2020 / Accepted on May 31, 2022

ABSTRACT. The aim of this paper is to analyse the evolution of the COVID-19 pandemic in Rio Grande do Sul by applying graph-theoretical tools, particularly spectral clustering techniques, on weighted graphs defined on the set of 167 municipalities in the state with population 10,000 or more, which are based on data provided by government agencies and other sources. To respond to this outbreak, the state has adopted a system by which pre-determined regions are assigned flags on a weekly basis, and different measures go into effect according to the flag assigned. Our results suggest that considering a flexible approach to the regions themselves might be a useful additional tool to give more leeway to cities with lower incidence rates, while keeping the focus on public safety. Moreover, simulations show that the combination of pendulum migration and isolation data used in this paper leads to a coherent qualitative description of the evolution of the pandemic in Rio Grande do Sul. These simulations also confirm the dampening effect of isolation on the dissemination of the disease.

Keywords: Spectral clustering, COVID-19 pandemic, discrete epidemiological model.

1 INTRODUCTION

The aim of this paper is to employ graph-theoretical tools to understand the dissemination of COVID-19 in the Brazilian state of Rio Grande do Sul. These tools may be useful sources of additional information for decision making by health and government authorities.

The year 2020 has been marked by the global outbreak and spread of the virus SARS-CoV-2, which causes the coronavirus disease (COVID-19) in humans [8]. In December 2019, several patients with an unknown severe respiratory disease were traced back to a wholesale market in

*Corresponding author: Luiz Emilio Allem – E-mail: emilio.allem@ufrgs.br
[1]Instituto de Matemática, Universidade Federal do Rio Grande do Sul, R. Bento Gonçalves, 9500, 91509-900, Porto Alegre, RS, Brasil – E-mail: emilio.allem@ufrgs.br https://orcid.org/0000-0001-9866-1541
Instituto de Matemática, Universidade Federal do Rio Grande do Sul, R. Bento Gonçalves, 9500, 91509-900, Porto Alegre, RS, Brasil – E-mail: choppen@ufrgs.br https://orcid.org/0000-0002-7581-1583
[3]Instituto de Matemática, Universidade Federal do Rio Grande do Sul, R. Bento Gonçalves, 9500, 91509-900, Porto Alegre, RS, Brasil – E-mail: matheus.marzo@ufrgs.br https://orcid.org/0000-0003-3390-6426
[4]Instituto de Matemática, Universidade Federal do Rio Grande do Sul, R. Bento Gonçalves, 9500, 91509-900, Porto Alegre, RS, Brasil – E-mail: lucas.siviero@ufrgs.br https://orcid.org/0000-0003-3347-5064

Wuhan, China. Researchers were quick to detect and isolate a novel strain of coronavirus [26]. It was soon discovered that the virus is highly contagious, and that it can be transmitted by infected individuals before they show the first symptoms and even by infected individuals that remain asymptomatic throughout the course of the disease [25]. This has led to unprecedented public health measures by the Chinese authorities. A lockdown of Wuhan and 15 other cities in Hubei Province took effect on January 23 [11]. On January 30, the World Health Organization (WHO) declared COVID-19 a public health emergency of international concern [8]. In the next month, a large number countries implemented measures aiming to prevent a global pandemic, ranging from travel restrictions, contact tracing and social isolation to border closures and lockdowns [8]. These actions turned out to be unsuccessful in eradicating the disease, and the WHO characterised the outbreak as a pandemic on March 11. Two days later, it assessed that Europe had become the epicenter of the pandemic [8]. The virus then quickly reached Brazil.

The first confirmed case of COVID-19 in Brazil dates back to February 26, in the state of São Paulo, and the Brazilian Health Ministry declared a state of nationwide community transmission on March 20 [3]. At that point, the number of confirmed cases in the state of Rio Grande do Sul was 37 [6], and the state government had already instated measures aimed at slowing down the spread of the virus, including school closures and a ban on commercial interstate travel [4]. In the next month, a large number of restrictions were imposed on activities that were deemed inessential. By the end of April, there were 1466 cases and 51 deaths officially attributed to COVID-19 in the state of Rio Grande do Sul [6]. At this point, and as of this writing, there was no vaccine or proven effective treatment for patients with severe cases of COVID-19 [8]. Recognizing the seriousness of the health crisis and the social and economic impact of widespread isolation, the state government unveiled a regulatory model for *controlled distancing* [5][1], which went into effect on May 11.

This regulatory model divided the state into 20 (pre-determined) regions based on the availability of beds in intensive care units (ICU beds) for COVID-19 patients. Every week, each region is assigned one of four possible *flags*, yellow (low risk), orange (medium risk), red (high risk) or black (very high risk), according to a numerical value based on several indices that measure the spread of the disease and the availability of ICU beds. Each flag entails different social distancing measures and imposes different constraints on businesses (or even their mandatory closure). This regulation has legal precedence over more flexible measures determined by local authorities or by the federal government [2]. Due to its effect on daily lives and on the economic activity, this model has been in the spotlight, and it has mustered praise, but also faced criticism. We should mention that, after the adoption of this system in Rio Grande do Sul, other states have followed suit and devised similar models (for instance, Acre, Mato Grosso, Mato Grosso do Sul, Pará, Rio de Janeiro, and São Paulo).

The general aim of this paper is to analyse the evolution of the COVID-19 pandemic in Rio Grande do Sul by applying graph-theoretical tools on data provided by government agencies and other sources. Given the flag system described above, we believe that clustering techniques are

[1] `https://distanciamentocontrolado.rs.gov.br/`

particularly well-suited for this analysis. The general idea of clustering is to partition a (typically large) data set into (a much smaller number of) clusters in a way that data in a same cluster are *similar*, and data in different clusters are *dissimilar*. Formal measures of affinity and of the quality of a given partition rely heavily on the context of the problem being considered. In this paper, we address clustering from a graph-theoretical perspective. We consider weighted graphs $G = (V, E, \omega)$, where the *vertex set* V is the data set, the edge set E contains edges connecting elements of V and the function $\omega\colon E \to \mathbb{R}_{>0}$ assigns a positive weight w_{ij} to each edge $ij \in E$. Our main tool is the use of spectral clustering, which is widely used in exploratory data analysis, but, to the best of our knowledge, has not been explored in connection with epidemiological models. Specifically, we would like to contribute in the following directions:

(a) So far, social distancing is the only measure to contain the spread of the disease. What would be a sensible way of dividing the state into smaller regions, so that each city lies in a cluster with cities to which it is strongly connected? How does this division relate with geographical divisions used by the state government? Is this interconnection reflected in the manner in which the disease has actually spread in the state? Did the response to self-isolation measures affect the way in which cities were interconnected?

(b) The flag system proposed by the government prescribes constraints on activities and businesses according to the risk assigned to each region. Would a more flexible approach, in which cities can be assigned to different regions on a weekly basis, allow more cities to be assigned lower risk flags?

(c) The availability of data, and the quality of the data, is fundamental to get meaningful results from any mathematical model. Did our data accurately capture the movement between cities and rates of social isolation? Can we see the impact of social isolation on the dissemination of the disease?

To address the first two questions, we consider two types of affinity measures. The first type is based on pendulum migration between cities, by which we mean the daily flow of commuters for work or education, to which we incorporate data about self-isolation. Our method gives a partition based solely on pre-pandemic data that captures the connection between the clusters and the spread of the disease. Moreover, we observed that incorporating isolation had a negligible effect on the way cities are clustered together. We believe that this suggests that the reaction to appeals for isolation was similar throughout the state, regardless of the particular way in which each city was affected by the disease. This seems to highlight the importance of a coordinated message by federal, state and local, and by the media. The second type of affinity measures is based on the availability of ICU beds. In this case, considering a more flexible approach to the regions, by which new clusters are determined on a weekly basis, more cities are assigned lower-risk flags. This may be useful complementary information for the flag system used in Rio Grande do Sul.

As a means to assess the quality of our data, we have also used a discrete SEIR compartmental model to simulate the spread of the disease and the effect of the social distancing measures that have been implemented, based on the migration and isolation data used for clustering. In contrast to clustering techniques, models of this type are a basic tool in the epidemiological toolbox, both in their discrete and continuous versions, and there is a vast literature related with them, see [13, 14] and the references therein. Our contribution in this respect was to show that the data for pendulum migration and isolation, combined with the available disease information, described a scenario that is coherent with the evolution of the disease in the state. Extrapolating from this, we conclude that isolation measures have been very important in slowing down the spread of the disease (often referred to as *flattening the curve* of new cases).

The remainder of the paper is organized as follows. In Section 2, we describe the data used in this paper. Section 3 is concerned with spectral clustering and its mathematical foundations. The affinity measures mentioned above are discussed in that section, and we also analyse the partitions that have been obtained by spectral methods. The SEIR model is introduced and analysed in Section 4. We finish the paper with concluding remarks.

2 DATA

In this section, we describe the data used in our study. The actual matrices are available in our git repository[2]. We consider the 167 municipalities in the state of Rio Grande do Sul whose estimated population in 2019 is above 10.000 according to the Brazilian Institute of Geography and Statistics (IBGE) [3]. Hereafter they will be referred to as *cities*. The distance between cities is given by a square matrix $\mathscr{D} = (d_{ij})$ of order $n = 167$, where d_{ij} denotes the average road distance from the seat of municipality i to the seat of municipality j and vice-versa, as calculated by the web mapping service Google Maps.

Using data from the population census of 2010, which is the most recent census performed in Brazil, we define square matrices $\mathscr{T} = (t_{ij})$ and $\mathscr{E} = (e_{ij})$ of order n, where t_{ij} is the number of daily commuters who reside in i and work in j and e_{ij} is the number of commuters who reside in i and go to school in j. These matrices have been obtained by extracting anonymized census microdata related to long-form questionnaires, which are publicly available[4], and by extrapolating them to the entire city population (adjusted to the 2019 values) using the survey weights that are part of the census microdata. To extract the data from this large dataset, we used the commercial statistical software Stata..

We also considered data directly related to the spread of the disease, and to the response to it, which has been extracted directly from the state health authorities [6].

[2] https://www.github.com/Lucassib/Cluster-COVID-19-RS

[3] https://www.ibge.gov.br/estatisticas/sociais/populacao

[4] https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?=&t=downloads

In our approach, the *time* t is measured in weeks, where our weeks correspond to the state's *epidemiological weeks*, which go from Saturday to Friday. Regarding epidemiological data, we consider $N = 17$ weeks starting at the week of March 7-13, when the first cases of COVID-19 were officially confirmed in the state, until July 3. We note that most pandemic related data is actually released on a daily basis, but contains fluctuations that may be attributed to administrative procedures. For instance, the number of reported cases and deaths regularly goes down on weekends and holidays, and surges in the first business days thereafter, which suggests that it does not reflect the actual behavior of the disease. Regarding cases and deaths, the weekly data that we collect is simply the overall number of reported cases in a week. Regarding self-isolation and ICU beds occupancy rates, we take the average over the time period. We should point out that the number of ICU beds in the state expanded considerably during the weeks considered, so that the number of total ICU beds in each city is also tracked on a weekly basis. Finally, we point out that these data are only used for $N = 8$ weeks, starting at the week between May 2 and 8 (when the model for controlled distancing was unveiled).

The information about self-isolation in each city $i \in [n] = \{1, \ldots, n\}$ is given by values $\beta_i(t) \in [0,1]$ for all $t \in \{1,2,\ldots,N\}$. This is an index developed by In Loco[5], a technology firm with offices in Brazil and in the United States, calculated from granular anonymized geolocation data from more than 60 million mobile devices across Brazil. It is defined as the proportion of devices in a city i that stayed within a radius of 450 meters from their habitual home during day t [10,21].

3 CLUSTERING

Consider a set of points $M = \{p_1, \ldots, p_n\}$ such that a weight $w_{ij} \geq 0$ is assigned to each pair of points p_i and p_j, where $i \neq j$ and $i, j \in [n]$. The aim of data clustering is to partition this set of points into classes such that elements of the same class are more alike, while elements of different classes are less alike. The weight w_{ij} measures *affinity* or *similarity* in this context[6]; the larger the value, the larger their affinity. For general data sets, a large number of similarity measures appear in the literature, and their quality depends on the context in which they are used [23].

Here, points are cities and weights are used to measure whether cities are highly interconnected or not. Several such measures will be considered here. For instance, a simple way to measure interconnection between cities is by simply considering the number of people who commute between them. This leads to the following matrix, where the weight α_{ij} between cities i and j is defined through the matrices $\mathscr{T}$ and $\mathscr{E}$ defined in Section 2:

$$A_0 = (\alpha_{ij}), \text{ where } \alpha_{ij} = t_{ij} + t_{ji} + e_{ij} + e_{ji}. \tag{3.1}$$

This choice is justified because, in the context of affinity measures, it is natural to consider symmetric weights.

[5] `https://www.inloco.com.br/`

[6] We use the word affinity because, in some of our examples, cities that are more different in some aspects will have more affinity to each other.

In order to understand how the interconnection between cities was affected during the pandemic, we also considered weights given by matrices $A(t)$, for $t \in \{1,\dots,N\}$. To incorporate self-isolation data, we first adjust the rate of self-isolation in each city i in terms of the average isolation $\overline{\beta}_i$, which was calculated using the same cell-phone data for i in the entire month of February, before the implementation of measures to contain the dissemination of COVID-19. We define

$$\beta_i^*(t) = \max\left\{\frac{\beta_i(t)-\overline{\beta}_i}{1-\overline{\beta}_i}, 0\right\}, \tag{3.2}$$

so that $\beta_i^*(t) = 0$ if rates of self-isolation are below average (this actually does not happen in our data set after the first week); otherwise, it is a linear interpolation where 0 corresponds to the average rate and 1 to full isolation. We are now ready to define

$$A(t) = (a_{ij}), \text{ where } a_{ij} = \left(1-\beta_j^*(t)\right)(t_{ij}+e_{ij}) + (1-\beta_i^*(t))(t_{ji}+e_{ji}). \tag{3.3}$$

The definition of $A(t)$ reflects our belief that it is conceptually more relevant to consider information about isolation in city j to assess the impact on commuting from i to j than information about isolation in city i. On the other hand, we understand that the nature of our isolation index, which estimates the number of individuals who never leave their home, could suggest using indices in city i to limit commutes from i to j. This has been tested and would have negligible impact on the results. Moreover, it would have been natural to ignore data related to student mobility as of the third week because all in-person school and university operations had already been suspended by then. However, this turned out to make clustering more unstable, perhaps because entries associated with smaller or more remote cities became too small.

3.1 Normalized cut

Before introducing the other affinity measures used in this paper, we first describe the framework of our analysis. We think of the data points as vertices in a graph $G = (V,E)$, where we use $V = [n]$ for simplicity. The weight between p_i and p_j is viewed as a weight $\omega(ij) = w_{ij}$ associated with the edge ij of G (if $w_{ij} = 0$, we assume that vertices i and j are not adjacent in G).

In general terms, a clustering problem in $G = (V,E)$ consists of finding a partition $V = V_1 \cup \cdots \cup V_k$ of the vertex set into a pre-determined number k of classes, where the partition optimizes some measure of quality of the partition. There are several such measures proposed in the literature [23]. In this paper, we work with the the normalized cut introduced by Shi and Malik in [12]. To define it, some additional notation is needed. Given $U \subset V$, let $\overline{U} = V \setminus U$ be the complement of U with respect to V. Moreover, for $S,T \subset V$, let $W(S,T) = \sum_{i\in S, j\in T} w_{ij}$. For a partition $\mathscr{P} = \{V_1,\dots,V_k\}$ of V, let

$$\mathrm{NCut}(\mathscr{P}) = \sum_{\ell=1}^{k} \frac{\mathrm{Cut}(V_\ell, \overline{V_\ell})}{\mathrm{Vol}(V_\ell)}, \tag{3.4}$$

where

$$\mathrm{Cut}(\mathscr{P}) = \frac{1}{2}\sum_{\ell=1}^{k} W(V_\ell, \overline{V_\ell}) \text{ and } \mathrm{Vol}(V_\ell) = \sum_{i\in V_\ell}\sum_{j\in V} w_{ij}.$$

Finding an optimal partition in this context is to find a partition $\mathscr{P}$ of V that minimizes the value of $\mathrm{NCut}(\mathscr{P})$. Note that this objective function takes both aims of clustering into account. On the one hand, the only weights that appear on numerators of terms in (3.4) are weights of edges whose endpoints lie in distinct classes, so that minimizing the function favors partitions such that vertices in different classes have small weight. On the other hand, the denominator of the term associated with V_i in (3.4) counts the weight of each edge with both endpoints in V_i twice, while the other edges incident with V_i are only counted once. So, increasing the weight of internal edges would decrease the value of the cut. Unfortunately, the authors of [12] showed that the problem of finding such a partition is NP-hard for general graphs (even if $k = 2$).

However, this problem is well-suited for a spectral approach. The following definitions are well known in spectral graph theory. The *weighted adjacency matrix* $W = (w_{ij})$ of a graph $G = (V,E)$ with weight function ω is defined by $w_{ij} = \omega(ij)$ if $ij \in E$ and $w_{ij} = 0$ otherwise. The degree of a vertex $i \in V$ in G is given by $d_i = \sum_{j=1}^{n} w_{ij}$. The diagonal matrix with the degrees $d_1, \ldots, d_n$ on the diagonal is called the *degree matrix* D.

At this point, we could simply present the procedure that we use to cluster our data; however, we believe that explaining how it works, and its connection to linear algebra, clarifies our approach. The following computation are performed in detail in [23]. Given a positive integers n and k and a partition $\mathscr{P} = \{V_1, \ldots, V_k\}$ of the vertex set of a graph $G = (V,E)$ with weight function ω and no isolated vertices, consider the matrix $X_{\mathscr{P}} \in \mathbb{R}^{n \times k}$ whose columns are the k vectors $\mathbf{x}^{(\ell)} = ({x_1}^{(\ell)}, {x_2}^{(\ell)}, \ldots, {x_n}^{(\ell)})^T$ with coordinates

$$x_j^{(\ell)} = \begin{cases} \frac{1}{\mathrm{Vol}(V_\ell)} & \text{if } j \in V_\ell; \\ 0 & \text{otherwise,} \end{cases}$$

for all $\ell \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n\}$. Using the *Laplacian matrix* $L = D - W$ associated with the weighted graph G, it turns out that

$$\mathrm{NCut}(\mathscr{P}) = \sum_{\ell=1}^{k} \frac{\mathrm{Cut}(V_\ell, \overline{V_\ell})}{\mathrm{Vol}(V_\ell)} = \sum_{\ell=1}^{k} {\mathbf{x}^{(\ell)}}^T L \mathbf{x}^{(\ell)} = \mathrm{tr}(X_{\mathscr{P}}^T L X_{\mathscr{P}}).$$

Writing $Y_{\mathscr{P}} = D^{-\frac{1}{2}} X_{\mathscr{P}}$ we obtain that

$$\mathrm{NCut}(\mathscr{P}) = \mathrm{tr}(Y_{\mathscr{P}}^T (D^{-\frac{1}{2}} L D^{-\frac{1}{2}}) Y_{\mathscr{P}}) = \mathrm{tr}(Y_{\mathscr{P}}^T \mathscr{L} Y_{\mathscr{P}}),$$

where $\mathscr{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$ is the *normalized Laplacian matrix* associated with G. Therefore finding an optimal partition in the sense of [12] is equivalent to finding a partition $\mathscr{P}$ that minimizes

$$\mathrm{ncut}_k(G) = \min_{\mathscr{Q}} \mathrm{NCut}(\mathscr{Q}) = \min_{\mathscr{Q}} \mathrm{tr}(Y_{\mathscr{Q}}^T \mathscr{L} Y_{\mathscr{Q}}),$$

where $\mathscr{Q}$ ranges over all partitions of V into exactly k sets. It is easy to see that $Y_{\mathscr{Q}}^T Y_{\mathscr{Q}} = I$, and by the Rayleigh-Ritz Theorem [17, Theorem 13], we have

$$\min_{Y \in \mathbb{R}^{n \times k}, Y^T Y = I} \mathrm{tr}(Y^T \mathscr{L} Y) = \lambda_1 + \cdots + \lambda_k, \tag{3.5}$$

where $0 = \lambda_1 \leq \cdots \leq \lambda_k$ are the k smallest eigenvalues of the symmetric matrix $\mathscr{L}$. Moreover, equality is achieved by matrices Y whose columns are orthogonal unit vectors generated by eigenvectors associated with the eigenvalues $\lambda_1, \ldots, \lambda_k$. As we have discussed, each partition of V into k parts is associated with a matrix Y as above. However, there are matrices Y that are feasible for (3.5), but are not of the form $Y_{\mathscr{Q}}$ for any partition $\mathscr{Q}$. This leads to the the following inequality:

$$\operatorname{ncut}_k^{rel}(G) = \min_{Y \in \mathbb{R}^{n\times k}, Y^T Y = I} \operatorname{tr}(Y^T \mathscr{L} Y) \leq \operatorname{ncut}_k(G). \tag{3.6}$$

As in usual LP-relaxations, the left-hand side of the inequality (3.6) may be computed efficiently and gives a lower bound on the value of an optimal partition. On the other hand, there is no obvious connection between a matrix Y that achieves $\operatorname{ncut}_k^{rel}(G)$ in (3.6) (i.e. a matrix constructed from eigenvectors associated with the smallest eigenvalues of $\mathscr{L}$) and a partition into k parts $\mathscr{P}$ such that $\operatorname{NCut}(\mathscr{P})$ is close to $\operatorname{ncut}_k(G)$. The following heuristic tries to find good quality partitions. To turn the matrix Y into a partition $\mathscr{P}$, it uses a well-known geometric method, known as K-means [16]. One way of assessing the quality of the output partition $\mathscr{P}$ is by looking at the ratio $\operatorname{NCut}(\mathscr{P})/\operatorname{ncut}_k^{rel}(G) \geq 1$. If this ratio is exactly 1, the partition $\mathscr{P}$ is optimal. Otherwise, it gives an upper bound on the actual value of the ratio $\rho(\mathscr{P}) = \operatorname{NCut}(\mathscr{P})/\operatorname{ncut}_k(G)$ (however, we should mention that the gap between $\operatorname{ncut}_k^{rel}(G)$ and $\operatorname{ncut}_k(G)$ may be very large in general). It is important to mention that this heuristic has been quite successful in practice, we refer to [18, 19, 24] for more explanation about these empirical findings. Moreover, defining the best choice for the number of clusters k is an important problem with no definitive solution. Parameters that are often used to indicate a good choice of k are the spectral gap (this is the ratio between consecutive eigenvalues, small ratios followed by a larger jump λ_{k+1}/λ_k indicate that k is a good choice) and the closeness to 0 (k is the number of eigenvalues below a certain threshold), and the stability of the clusters obtained in repeated iterations of the procedure, but other criteria also appear in the literature [23].

We now state the heuristic of Shi and Malik [12], iterated S times. Given an affinity matrix W associated with an n-vertex graph $G = (V,E)$, do the following:

(1) Let D to be the degree matrix associated with W and construct its normalized Laplacian matrix $\mathscr{L} = D^{-1/2} L D^{-1/2}$, where $L = D - W$.

(2) Compute vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, where each $\mathbf{x}_i$ is a unit eigenvector associated with the eigenvalue λ_i, where $\lambda_1, \ldots, \lambda_k$ are the k smallest eigenvectors of $\mathscr{L}$ (counting multiplicity). In the case of repeated eigenvalues, the eigenvectors associated with them must be orthogonal. Form the matrix $X = [\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_k] \in \mathbb{R}^{n\times k}$ by stacking these eigenvectors in columns.

(3) Form the matrix $Y = (y_{ij})$ from $X = (x_{ij})$ by renormalizing each of the rows to have unit length (i.e. $y_{ij} = x_{ij} / \sum_{j=1}^{n} x_{ij}$).

(4) for $s = 1, \ldots, S$ do (let $\mathscr{Q}$ denote the best partition obtained up to a given step, where the starting partition is arbitrary.)

(4.1) Treating the ith row of Y as a point $\mathbf{y}_i \in \mathbb{R}^k$, split $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ into k clusters $S_1, \ldots, S_k$ via K-means.

(4.2) Let $\mathscr{P}$ be the partition such that vertex i is assigned to cluster ℓ if and only if $\mathbf{y}_i$ lies in S_ℓ.

(4.3) If $\mathrm{NCut}(\mathscr{P}) < \mathrm{NCut}(\mathscr{Q})$, redefine $\mathscr{Q}$ as $\mathscr{P}$.

(5) Return $\mathscr{Q}$, the partition with minimum Ncut obtained in step (4).

3.2 Affinity based on pendular migration

When we compute the eigenvalues of the matrix $\mathscr{L}$ associated with the affinity measure A_0 defined in (3.1), we find determine that there is a considerable eigenvalue gap between λ_{10} and λ_{11}, which suggests that $k = 10$ is a good choice for the number of clusters. When we apply the above procedure to the affinity measure A_0 for $S = 500$, we obtain the partition given in Figure 1, whose gap is $\mathrm{NCut}(\mathscr{P})/\mathrm{ncut}_k^{rel}(G) \approx 1.3256$. This means that $\mathscr{P}$ is at most 32.56% above the actual value of $\mathrm{ncut}_k(G)$, but the gap is typically much smaller (and may possibly be optimal). Regarding stability, this partition $\mathscr{P}$ has been obtained 183 times out of the 500 iterations of the procedure.

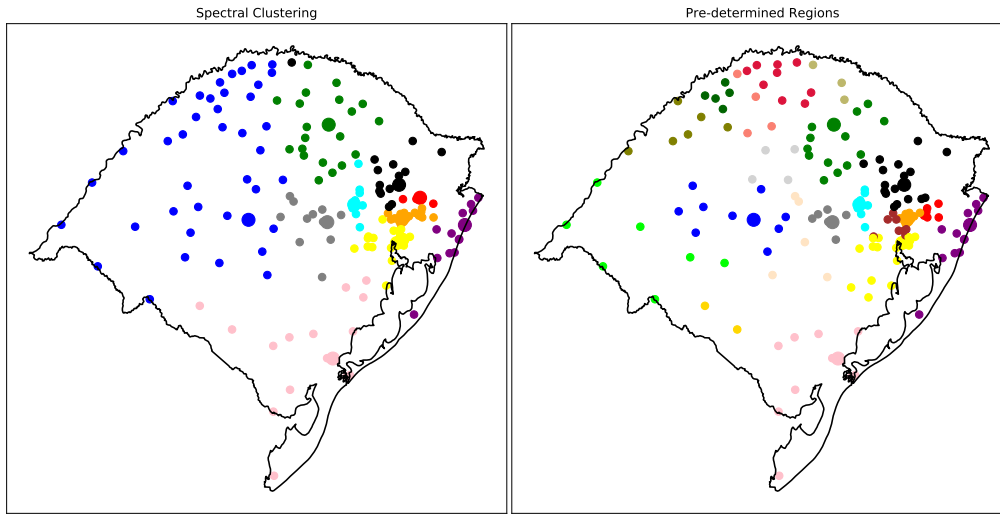

Figure 1: Clustering obtained by spectral clustering with respect to measure A_0 for $k = 10$ clusters. The largest city in each cluster is marked with a larger circle. The regions defined by the government are on the right.

Even though the data used to obtain this partition is not related to the pandemic, if we look at the evolution of the number of cases during this time period, the connection between the clusters and the spread of the disease is perceptible. For instance, Figure 2 shows how real data about

the disease evolved in cities of two neighboring clusters of Figure 1 (left), namely the black and red clusters, in four different weeks (detailed material for all clusters is available in our git repository). The red cluster consists of four cities: Gramado, Canela, Nova Petrópolis and São Francisco de Paula (which are part of a nationally renowned touristic area) and the other cluster is centered in Caxias do Sul, the second largest city in the state by population. One important feature about these clusters is that they are subsets of the same region, according to the state 20 pre-determined regions. Note that the largest cities of all remaining clusters are also the largest city in their pre-determined region. The first cases appear in the cluster of Caxias do Sul quite early, and they quickly spread to cities in the same cluster, which has a relatively large number of active cases by May 2, the first week displayed in the figure (and the ninth week with cases in the state). On the other hand, there are no recorded cases in the cluster of Gramado until the week of May 9. After the first case is identified, all the other cities in the cluster record cases in a span of three weeks.

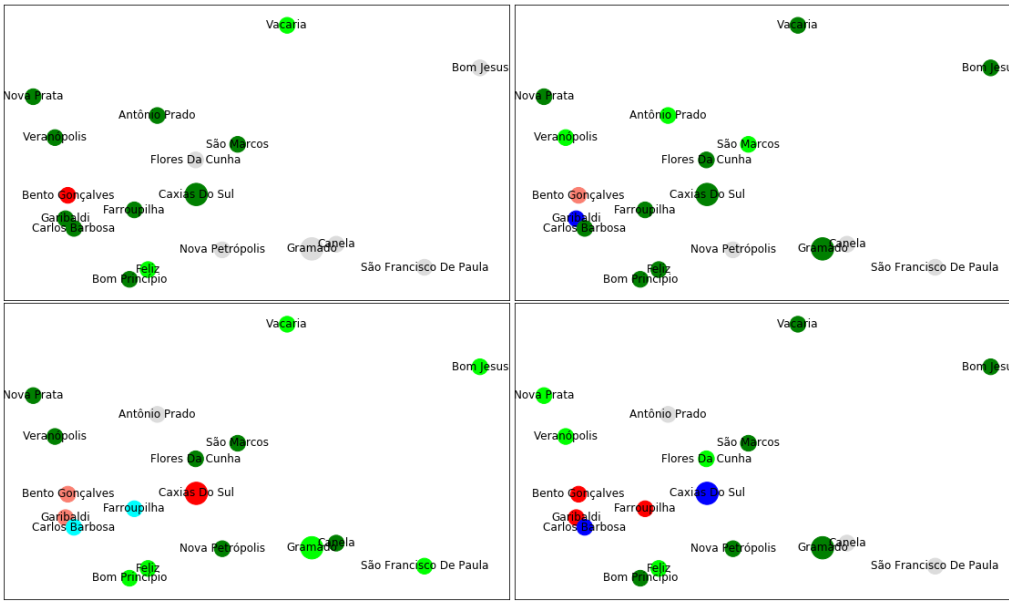

Figure 2: Clockwise, starting from the top left. Cases on the weeks from May 2-8, May 9-15, May 16-22 and from May 30 to June 5. Gray stands for no active cases, green for cases in the interval $[1,50]$, blue for $[51,100]$ and red above 100. Dark colors mean that the number of active cases has increased from the previous week, light colors mean that they have decreased.

This behavior supports the choice of pendulum migration as a footprint for the spread of the disease, as was done in [22], for instance. However, instead of census data, the authors of [22] used mobile geolocation data from [21] to monitor the movement between cities.

As mentioned at the beginning of this section, instead of using A_0, one could adjust the measure to incorporate rates of isolation, using a different measure $A(t)$ (defined in (3.3)) at each time

t. As it turns out, the difference in the partition obtained when performing the above clustering procedure for $A(t)$ instead of A_0 is minor. Indeed, the Hamming distance at any time t between the two partitions was at most 1 (out of 167). This may indicate that public response to self-isolation has been rather uniform throughout the state.

3.3 Affinity based on available ICU beds

As mentioned in the introduction, the state government introduced regulation to define when mandatory protocols of social distancing must be put into effect. Every Saturday, each region, out of a pre-determined set of 20 regions (which in turn are sorted into seven *macroregions*), is assigned one of four possible flags, yellow, orange, red or black, according to a numerical value based on several indices, which take the number of cases, the number of hospitalizations, the number of deaths and the availability of ICU beds into account. Once a flag has been assigned, cities in the region must adapt to the state regulations associated with that flag (local governments may enforce stricter rules, if desired).

Even though this method was met by a very positive reception from health and local authorities, its implementation quickly led to complaints by cities and economic agents who deem to have been treated unfairly. For instance, in the first weeks using this method, it was pointed out that several cities where no cases had ever been recorded had been assigned orange or red flags (owing to an outbreak or a shortage of ICU beds in their region, for instance). Moreover, since the index for a region incorporates data from the macroregion to which it belongs, a high risk flag can be assigned to a region in which no city had a substantial number of cases. In some instances, this has led to loud public outcry and threats of disobedience by local authorities, which in turn led to negotiations and adjustments. At the present moment, regulations include automatic 'flag reductions' for cities that meet certain criteria. This is the case for cities where no new cases have been recorded in the past two weeks, for instance. Moreover, each city can appeal to a board after its weekly classification has been revealed. When this happens, the city is allowed to present new data, such as an expansion on the total number of ICU beds.

Given this reality, we aim to look at the partition into regions under a more flexible perspective. To this end, we propose affinity measures that consider the availability of ICU beds (updating it weekly) and consider what happens when we re-organize the regions on a weekly basis. For a city i, let $u_i(t)$ be the average total number of ICU beds in i at time t, and let $\ell_i(t)$ be the average number of ICU beds that are available (i.e. unoccupied and ready to accommodate new patients) in i at time t. The first measure is 'static', as it only considers the total number of ICU beds at the beginning of the recording process:

$$C_0 = (\gamma_{ij}), \text{ where } \gamma_{ij} = \frac{|u_i(0) - u_j(0)|}{d_{ij} + c}, \tag{3.7}$$

where $u_i(0)$ denotes the total number of ICU beds in city i on May 2 and d_{ij} is the distance between i and j given by matrix $\mathscr{D}$ (see Section 2) and the constant $c = 10$ avoids the effect of very small distances.

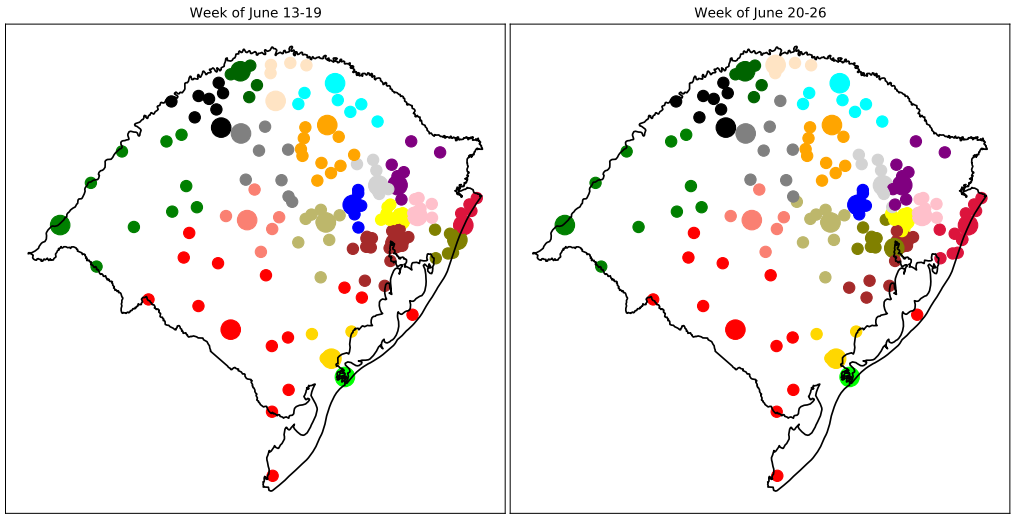

Figure 3: Partitions obtained using the affinity measure (3.8) using data from the weeks from June 13-19 (left) and June 20-26 (right).

The intuition behind this definition is that the health systems of two cities i and j that are geographically close, but whose health infrastructure is very different, would tend to be interconnected (with the city with small health capability transferring patients to the other), while two cities whose health capacities are equivalent would be less dependent on each other.

The second measure is 'dynamic', not only updating the number of ICU beds, but also considering the actual number of ICU beds that are ready to accommodate new patients:

$$C(t) = (c_{ij}(t)), \text{ where } c_{ij} = \max\left\{\eta_i(t)\frac{\ell_j(t)-\ell_i(t)}{d_{ij}+c}, \eta_j(t)\frac{\ell_i(t)-\ell_j(t)}{d_{ij}+c}\right\}, \tag{3.8}$$

where $c = 10$ and $\eta_i(t) = \frac{u_i(t)-\ell_i(t)+1}{u_i(t)+1}$. This quantity $\eta_i(t)$ may be viewed as a rate of urgency for city i to look for ICU beds outside its borders. This rate is 1 if it does not have any ICU beds or if all its ICU beds are occupied, and decreases as the percentage of available beds gets larger. The term $\ell_i(t) - \ell_j(t)$ accounts for the fact that a city j with more available ICU beds than i would be desirable to receive patients rom i. In other words, the affinity measure of interconnection between i and j goes up from the perspective of i if its health system is strained and j is geographically close and has more available beds.

Applying the above spectral partitioning procedure with the affinity measure defined in (3.8) for $k = 20$ (the number chosen by the state) and $S = 500$ produces the partitions in Figure 3 in two consecutive weeks. In this particular case, 26 cities switched regions from one week to the next.

Our aim using this measure is to assess whether allowing the regions to be re-organized on a weekly basis can bring meaningful additional information to one of the features of the state flag system, namely that the state consists of 20 pre-determined regions, which are in turn combined

into seven macroregions. To this end, we shall first give a general description of the way in which the state assigns flags to regions (the formula is in the appendix). The flag is based on 11 individual indices, classified in two main types, disease propagation or healthcare capacity, and computed in one of three levels (within each region, within each macroregion or statewide). For each index, four intervals have been defined, and a flag is assigned to the index according to the interval it belongs to. The flag actually assigned to the region is obtained from a weighted average of the flags assigned to the different indices.

Here, we have devised an alternative formula (the formula is in the appendix), which uses exactly the same indices wherever possible. An important difference is that we do not use any indices related with macroregions, as it would not make sense to assign a city to a new region every week, while at the same time assume that cities lie in a fixed macroregion. Unfortunately, some of the data available for macroregions was not publicly available, or was less reliable, for the cities themselves. Because of this, we transferred the weight of these indices to other indices measuring similar features for cities. To assess what the dynamic clustering obtained using our matrices might say about the clustering defined by the state, we proceed in two steps. The first compares the flags assigned to the 20 pre-determined regions using the state's formula and this new formula. Figure 4 does this for the weeks from June 13-19 and June 20-26. (A comparison for all seven weeks under consideration may be found in our git repository).

The second step is to split the state in 20 regions on a weekly basis (which we call the *dynamic partition*) and compare the flags assigned by the new formula to these regions and to the 20 pre-determined regions. This is done in Figure 5, which suggests that more cities would be assigned a lower-risk flag in the dynamic partition. On the week from June 13-19, 26 cities had a lower-risk flag for the dynamic regions, 13 cities had a higher-risk flag for the dynamic region, and 128 cities remained the same. On the week from June 20-26, these numbers where 50, 2 and 115, respectively.

In short, our computations suggest that the flags assigned with the new formula are related with the flags from the original formula. Moreover, flags assigned in the second step show that partitioning on a weekly basis allows for more flexibility than considering the same partition throughout. This sends the message that it might be possible to devise a formula that takes more, or more reliable, information into account (as in the government's formula), and that allows regions to be adapted on a weekly basis.

We should emphasize that we do not believe that the new formula presented here is better than the formula used by the state government, quite the opposite, but simulations suggest that the new formula was able to capture the main features of the government's formula using the data available to us. We are also not suggesting that our regions are necessarily better than the pre-determined regions defined by the state government. Even though our results show that a more flexible approach would allow more cities to be assigned lower-risk flags, implementing weekly changes to the regions would bring its own challenges. The government's regions are heavily based on the way in which the public health system is organized and on the reality that many cities of small and average size do not have hospitals, particularly hospitals equipped with ICU

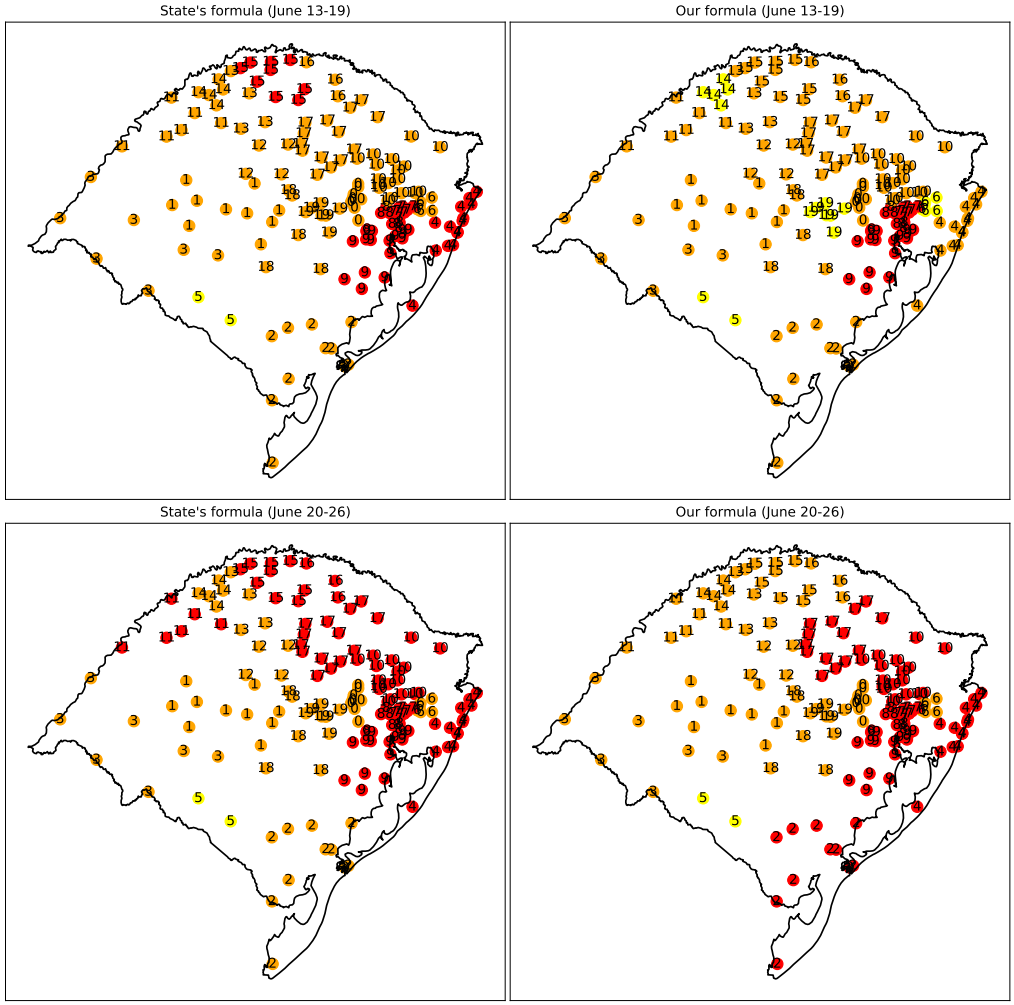

Figure 4: Flags assigned by the state formula (left) and by our formula (right) on the weeks from June 13-19 (top) and June 20-26 (bottom).

beds to treat complex cases, and therefore need to establish formal agreements with one or more cities to which their patients can be transferred. Because of this, periodic changes to the regions would require that some cities direct their patients to hospitals in different cities every week, which is certainly not easy to implement. However, in exceptional situations such as a pandemic, this might be justified, and accepted by local governments, given the benefit of more leeway to cities that are not as directly affected by the disease.

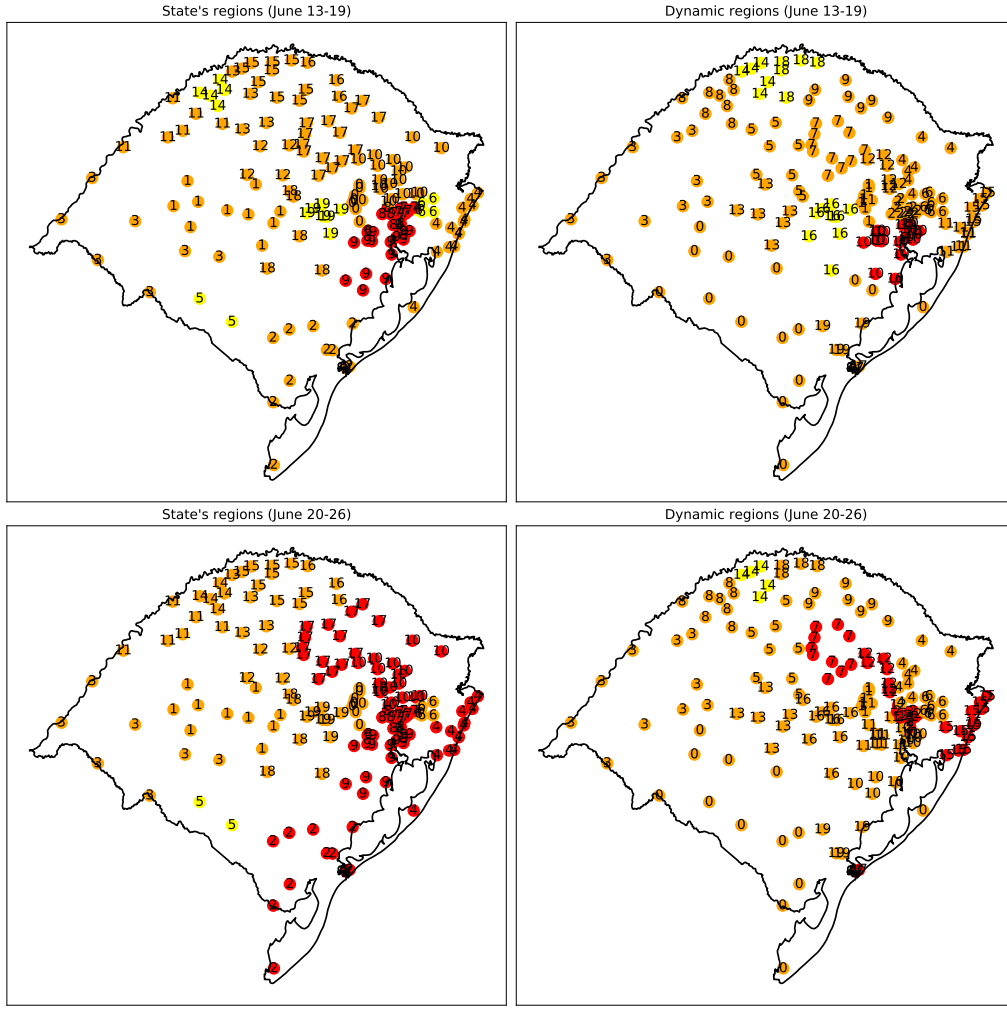

Figure 5: Flags assigned by our formula to the state's regions (left) and to the dynamic regions (right) on the weeks from June 13-19 (top) and June 20-26 (bottom). Flags are given by colors, regions are labelled 0 to 19.

4 SEIR MODEL

Using the data collected in the previous sections, it is possible define a discrete model for the spread of the disease, which gives a qualitative description of the evolution of the disease and helps us understand the effect of different parameters associated with the disease and of measures to contain it. We consider a discrete susceptible-exposed-infectious-recovered (SEIR) epidemiological model, where the spread of the disease is represented by a recurrence relation indexed by a discrete parameter $t \in \{0,1,\ldots\}$. This recurrence relations are give the expected behavior of a stochastic process defined on a digraph $G = (V,E,\omega)$, where each vertex represents a city and

the weight w_{ij} of an arc ij represents the number of commuters from i to j on an average day. Each city $i \in V$ has population P_i and, for all $t \geq 0$, the vector $\mathbf{x}_i(t) = (S_i(t), E_i(t), I_i(t), R_i(t))$ stands for the number of *susceptible*, *exposed*, *infected* and *removed* inhabitants of city i at time t, respectively. As usual, all susceptible individuals are assumed to be prone to contracting the disease. Exposed individuals have been infected, but are not yet contagious, while infected individuals are capable of infecting susceptible individuals. Removed individuals either recovered (and became immune from the disease) or passed away. Initially, each city i is assigned a vector $\mathbf{x}_i(0)$ with the number of individuals in each class at the start of the process.

We now describe how our system evolves. As in the work of Silva, Pereira and Nonato [22], we assume that most of the movement between cities may be attributed to daily commutes. On day t, part of the population of each city leaves their city to work or study, and comes back in the evening. This leads to a row stochastic matrix $M = (p_{ij})$ of order n, where $n = |V|$. We interpret p_{ij} as the *relative flow* from city i to city j, given by $p_{ij} = (t_{ij} + e_{ij})/P_i$, where t_{ij} and e_{ij} come from the matrices $\mathscr{T}$ and $\mathscr{E}$ from Section 2. This corresponds to the proportion of the population of i that regularly commutes to j. The diagonal entries are given by $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$.

As a consequence, during the day each city j has an effective population of

$$P'_j = \sum_{i \in V} p_{ij} P_i.$$

We shall also assume that all classes of individuals are equally likely to move between cities, so that the effective number of individuals of each class in city j on day t is given by

$$\begin{aligned} S'_j(t) &= \sum_{i \in V} p_{ij}(t) S_i(t),\ E'_j(t) = \sum_{i \in V} p_{ij}(t) E_i(t), \\ I'_j(t) &= \sum_{i \in V} p_{ij}(t) I_i(t),\ R'_j(t) = \sum_{i \in V} p_{ij}(t) R_i(t). \end{aligned}$$

In our model, infections only occur during the day (at the city where each individual spends the day). Each such individual is assumed to meet L other individuals in a normal day. However, assuming that a susceptible individual spends the day at city j, the number of actual meetings on day t is assumed to be $L(1 - \beta_j^*(t))^2$, where $\beta_j^*(t)$ is the relative rate of isolation of city j on day t, given in (3.2). This rate has been assumed under the simplifying assumption that the probability that, for a meeting to happen, both participants cannot be under self-isolation, and this would happen with probability $(1 - \beta_j^*(t))^2$ if the decision to self-isolate were taken by each individual spending the day in city j, independently of all others, with probability $\beta_j^*(t)$. When an individual is infected, we assume that the disease takes its course in 14 days, following the phases described in guidelines of the Center for Disease Control and Prevention (CDC) [9]. In the first four days [15], incubation occurs, in the next 5 days, infected individuals are contagious [20] and, in the final five days, individuals are still convalescent, but do not transmit the disease [7]. While infectious, we assume that the probability that an encounter between a susceptible and an infected individual leads to an infection is given by $\tau(k)$, where k is the number of days since the infected individual became contagious. As in [20], we assume that $\tau(k)$ follows a triangular

distribution over the five days, with a peak on the third day. We have $\tau(k) = 0$ for $k > 5$. The area of the triangle in the definition of this distribution is given by R_0/L, to ensure that the basic reproductive number (assuming no isolation) is $R_0 = 2.4$, following a situation report by the WHO [1] (see also [7]).

The recurrence relations become

$$\begin{aligned}
S_i(t+1) &= S_i(t) - R_0 S_i(t) \sum_j p_{ij} \frac{\sum_{k=t-4}^{t} (1-\beta_j^*(t))^2 I_j'^{new}(k)\tau(k-t+5)}{P_j'(t)} \\
I_i^{new}(t+1) &= R_0 S_i(t) \sum_j p_{ij} \frac{\sum_{k=t-4}^{t} (1-\beta_j^*(t))^2 I_j'^{new}(k)\tau(k-t+5)}{P_j'(t)} \\
E_i(t+1) &= E_i(t) + I_i^{new}(t+1) - I_i^{new}(t-2) \\
I_i(t+1) &= I_i(t) + I_i^{new}(t-2) - I_i^{new}(t-13) \\
R_i(t+1) &= R_i(t) + I_i^{new}(t-13)
\end{aligned}$$

In the above, for simplicity, we assume that $I_i^{new}(s) = 0$ for all $s \leq 0$ and $i \in [n]$. Just to illustrate where these equations come from, we discuss the case where an individual in city i does not contract the disease at time $t+1$ in the case where there is no social distancing. With probability p_{ij}, the individual moved to city j on day $t+1$. The probability that an encounter leads to an infection is

$$\frac{R_0}{L} \sum_{k=t-4}^{t} \tau(k-t+5) \frac{I_j'^{new}(k)}{P_j'(t)},$$

so that the probability that no encounter leads to an infection, given that the individual spends the day in city j, is

$$\left(1 - \frac{R_0}{L} \sum_{k=t-4}^{t} \tau(k-t+5) \frac{I_j'^{new}(k)}{P_j'(t)}\right)^L \approx 1 - R_0 \sum_{k=t-4}^{t} \tau(k-t+5) \frac{I_j'^{new}(k)}{P_j'(t)}.$$

Since the same holds for each susceptible individual in i and knowing the proportion of susceptible individuals that commute from i to j, the first equation in the above system gives the expected number of susceptible individuals at time t that remain susceptible at time $t+1$.

We run this model starting with the official state data on May 26 to simulate the evolution of the disease until July 9. The number of new infections in the days before this date are estimated using data from May 20-26, where we assume that new cases correspond to 10% of the number of active cases. The results for the cities of Porto Alegre (the state capital and largest city), Rio Grande (the largest port in Southern Brazil and the city with highest average rate of self-isolation) and Antônio Prado (a small city with a population of about 13,000, where the average rates of self-isolation are lowest) appear in Figure 6.

It is striking to compare it with the behavior of these quantities in the case where there is no social distancing (that is $\beta_j^*(t) = 0$ for all j and t) and with the situation in which the high rates of self-isolation observed on the week between March 21 and 27 had been maintained after May 26 ($\beta_i = 0.614$, on average). This appears in Figure 7.

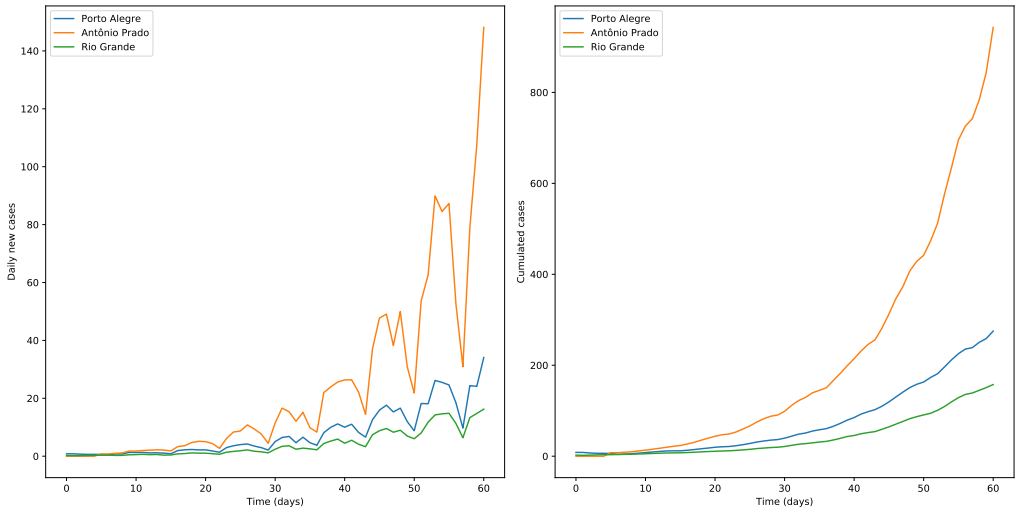

Figure 6: Number of new cases and the cumulative number of cases (per 100,000 inhabitants) in three cities of Rio Grande do Sul. The x axis represents the number of days after May 26.

To see the effect of self-isolation in this model, in Figure 8 we plot the number of cases in Porto Alegre on July 9 assuming that the rate of self-isolation remained constant throughout the time period, and is given by the corresponding value on the x-axis.

According to our data, the average rate of isolation in Porto Alegre has been about 44.3% during this time period. We note that a simple calculations shows that, while the number of susceptible individuals is much higher than the number of individuals in the other classes, isolation would need to be above 55% to keep the effective reproductive number of the disease below 1.

Even though we have opted to plot the evolution of the disease from May 26 to avoid intrinsic errors coming from initial conditions where the number of infected individuals was very small, we should mention that the isolation data were successful at explaining the ups-and-downs in the number of cases in the first weeks of the pandemic in Porto Alegre. According to the simulations, the number of cases remained stable between March 31 and the week of May 26, and started growing rapidly since then. State data report that the number was stable until early June, and grew rapidly since then. (Specific data are in the ancillary files.)

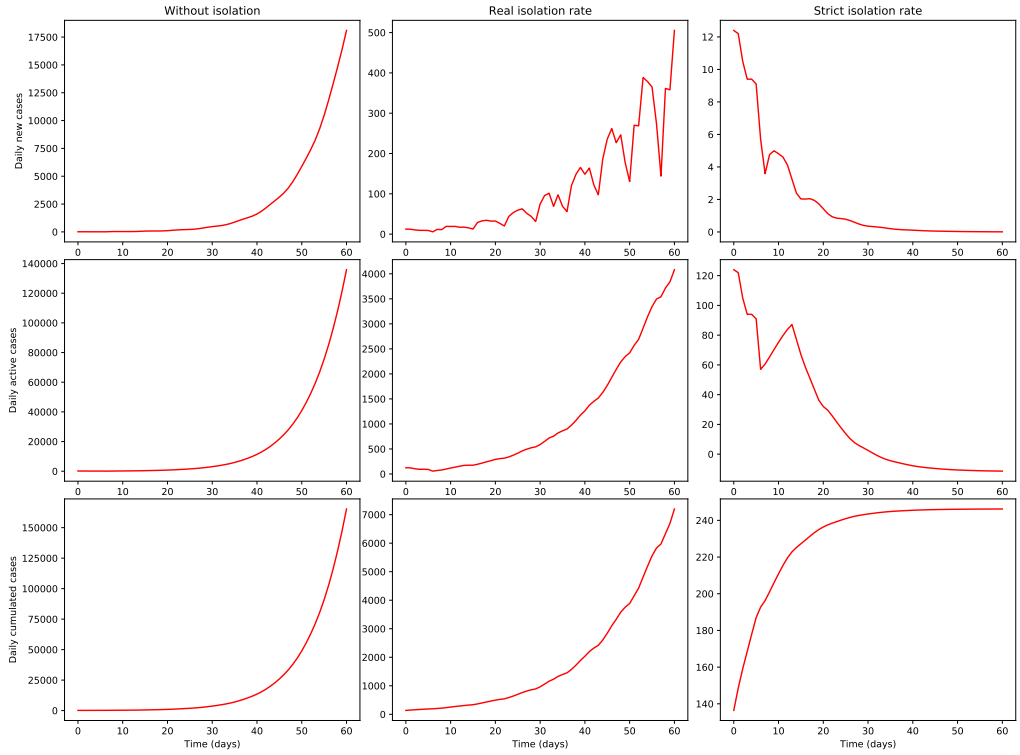

Figure 7: On the top: Number of new cases in Porto Alegre assuming the actual isolation data (left), no isolation (center) and strict isolation (right). In the middle: number of active cases in each scenario. On the bottom: cumulative number of cases in each scenario.

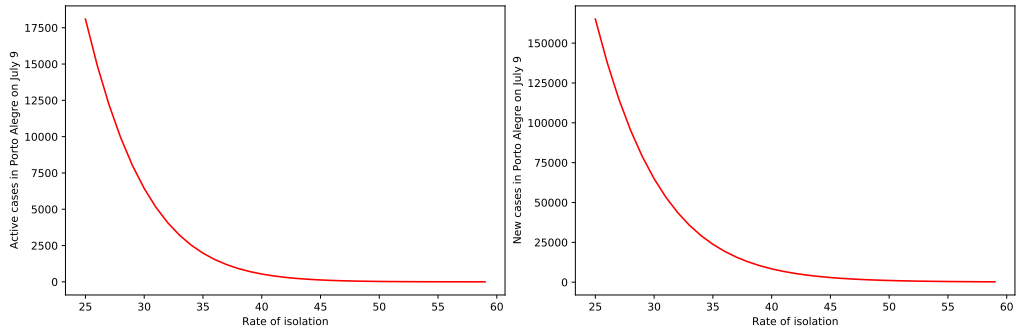

Figure 8: Number of active cases, and the cumulative number of cases in Porto Alegre on July 9, assuming that the rate of isolation remains constant, and is given by the value on the x axis.

5 CONCLUDING REMARKS

In this paper, we looked at the evolution of the COVID-19 pandemic in Rio Grande do Sul using graph theory. We applied spectral clustering techniques on weighted graphs defined on the set of 167 municipalities in the state with population 10,000 or more, using official data provided by government agencies and isolation data by In Loco. Results related with our first measure, based on data for pendulum migration, provided a partition of the state into 10 clusters. The largest city in all but one of the clusters is also the largest city in its own governmental region, and the only exception gives two regions where the evolution of the disease was quite different. This confirms that pendulum migration is an important means of spreading the disease. Our results have also shown that, in this situation, considering dynamic clusters that incorporate self-isolation data would give essentially the same clusters. In future work, it would be interesting to see if this can also be observed in other regions, particularly if they are more heterogeneous.

Given the specific context of the flag system in Rio Grande do Sul, our main contribution was obtained using an affinity measure based on the availability of ICU beds. Our results suggest that considering a flexible approach to the regions themselves would be a useful additional tool in giving more leeway to cities with lower incidence rates, while keeping the focus on public safety. However, this is just a first step in evaluating the adequacy of such an approach. Future work could look for more data (in a municipal level), which would allow a direct comparison with the government system. Moreover, implementing this approach on the ground would require state and local authorities to assess the practicality of periodic changes to the regions. For instance, this would need to be met with changes in patient transfer protocols.

To evaluate the quality of the data used for clustering, we have observed that disease information from the literature, combined with the isolation data, have provided a coherent qualitative description of the evolution of the pandemic in Rio Grande do Sul using a simple discrete SEIR model. Extrapolating from this, we conclude that isolation measures have been very important in slowing down the spread of the disease. Of course, better results would be achieved with a better understanding of the behavior of the disease and with a model that takes more information into account.

Acknowledgments

The authors are particularly indebted to In Loco for providing data about self-isolation in the cities of Rio Grande do Sul and to Prof. Márcia Barbian for sharing her data about availability and occupancy of ICU beds in the state. The authors also thank Alisson Matheus Fachini Soares, Guilherme Tadewald Varella and Lucas da Rocha Schwengber for helpful discussions leading to this paper. C. Hoppen acknowledges the support of CNPq 308054/2018-0 and FAPERGS 19/2551-0001727-8. L. E. Allem acknowledges the support of FAPERGS 21/2551-0002053-9. M. M. Marzo acknowledges the support of CAPES. L. S. Sibemberg acknowledges the support of CNPq. We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

REFERENCES

[1] Coronavirus disease 2019 (COVID-19) Situation Report 46, World Health Organization (2020).

[2] Diário Oficial da União, Decisão Ação Direta de Inconstitucionalidade 6343, 1 de junho de 2020, Supremo Tribunal Federal (2020).

[3] Diário Oficial da União, Portaria 454, 20 de março de 2020, Ministério da Saúde (2020).

[4] Diário Oficial do Estado do Rio Grande do Sul, Decreto 55.128, 19 de março de 2020 (2020).

[5] Diário Oficial do Estado do Rio Grande do Sul, Decreto 55.240, 10 de maio de 2020 (2020).

[6] Painel Coronavírus RS- Secretaria Estadual de Saúde (2020. Access on: 1 Ago. 2020). URL `https://ti.saude.rs.gov.br/covid19/`.

[7] Report 9: Impact of non-pharmaceutical interventions to reduce COVID-19 mortality and healthcare demand, Imperial College COVID-19 Response Team (2020. Access on: 1 Ago. 2020).

[8] Timeline of WHO's response to COVID-19. "`https://www.who.int/news-room/detail/29-06-2020-covidtimeline`" (2020. Access on: 1 Ago. 2020).

[9] B. Adhikari, L. Fischer, B. Greening, S. Jeon, E. Kahn, G. Kang, G. Rainisch, M. Meltzer & M. Washington. COVID19Surge: a manual to assist state and local public health officials and hospital administrators in estimating the impact of a novel coronavirus pandemic on hospital surge capacity (2020).

[10] N. Ajzenman, T. Cavalcanti & D. Da Mata. More Than Words: Leaders' Speech and Risky Behavior during a Pandemic. *SSRN*, (2020). doi:10.2139/ssrn.3582908. URL `https://ssrn.com/abstract=3582908`.

[11] D. Cyranoski. What China's coronavirus response can teach the rest of the world. *Nature*, **579**(7800) (2020), 479–480.

[12] Jianbo Shi & J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8) (2000), 888–905.

[13] M.Y. Li, H.L. Smith & L. Wang. Global dynamics of an SEIR epidemic model with vertical transmission. *SIAM Journal on Applied Mathematics*, **62**(1) (2001), 58–69.

[14] K. Linka, M. Peirlinck, F. Sahli Costabal & E. Kuhl. Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions. *Computer Methods in Biomechanics and Biomedical Engineering*, (2020), 1–8.

[15] S. Ma, J. Zhang, M. Zeng, Q. Yun, W. Guo, Y. Zheng, S. Zhao, M.H. Wang & Z. Yang. Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries. *medRxiv*, (2020). doi:10.1101/2020.03.21.20040329. URL `https://www.medrxiv.org/content/early/2020/03/24/2020.03.21.20040329`.

[16] J. Macqueen. Some methods for classification and analysis of multivariate observations. In "In 5-th Berkeley Symposium on Mathematical Statistics and Probability" (1967), p. 281–297.

[17] J. Magnus & H. Neudecker. "Matrix Differential Calculus with Applications in Statistics and Econometrics (Revised Edition)". John Wiley & Sons Ltd (1999).

[18] B. Nadler, S. Lafon, R.R. Coifman & I.G. Kevrekidis. Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. In "Proceedings of the 18th International Conference on Neural Information Processing Systems", NIPS'05. MIT Press, Cambridge, MA, USA (2005), p. 955–962.

[19] A.Y. Ng, M.I. Jordan & Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In "Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic". MIT Press, Cambridge, MA, USA (2001), p. 849–856.

[20] C.M. Peak, R. Kahn, Y.H. Grad, L.M. Childs, R. Li, M. Lipsitch & C.O. Buckee. Comparative Impact of Individual Quarantine vs. Active Monitoring of Contacts for the Mitigation of COVID-19: a modelling study. *medRxiv*, (2020).

[21] P.S. Peixoto, D.R. Marcondes, C.M. Peixoto, L. Queiroz, R. Gouveia, A. Delgado & S.M. Oliva. Potential dissemination of epidemics based on Brazilian mobile geolocation data. Part I: Population dynamics and future spreading of infection in the states of Sao Paulo and Rio de Janeiro during the pandemic of COVID-19. *medRxiv*, (2020). doi:10.1101/2020.04.07.20056739. URL `https://www.medrxiv.org/content/early/2020/04/11/2020.04.07.20056739`.

[22] P.J.S. Silva, T. Pereira & L.G. Nonato. Robot dance: a city-wise automatic control of Covid-19 mitigation levels. *medRxiv*, (2020). doi:10.1101/2020.05.11.20098541. URL `https://www.medrxiv.org/content/early/2020/05/18/2020.05.11.20098541`.

[23] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, **17**(4) (2007), 395–416.

[24] U. von Luxburg, M. Belkin & O. Bousquet. Consistency of spectral clustering. *Ann. Statist.*, **36**(2) (2008), 555–586. doi:10.1214/009053607000000640. URL `https://doi.org/10.1214/009053607000000640`.

[25] Y. Wang, J. Tong, Y. Qin, T. Xie, J. Li, J. Li, J. Xiang, Y. Cui, E.S. Higgs, J. Xiang & Y. He. Characterization of an asymptomatic cohort of SARS-COV-2 infected individuals outside of Wuhan, China. *Clinical Infectious Diseases*, (2020). doi:10.1093/cid/ciaa629. URL `https://doi.org/10.1093/cid/ciaa629`.

[26] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu & G. Gao. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, **382** (2020). doi:10.1056/NEJMoa2001017.

A FORMULA FOR FLAGS

This appendix contains the formulae used to compute the flags. The formulae are weighted means of a series of parameters.

A.1 Government formula

The parameters used in the government formula are as follows:

$R_j =$ Number of cities in region $j \in \{1,2,\ldots,20\}$.

$M_k =$ Number of cities in macroregion $k \in \{1,2,3,4,5,6,7\}$, where $M_{k(j)}$ is the macroregion containing region j.

$S =$ Total number o cities (167)

$P_j =$ Population of region j

$a(t) =$ New hospitalizations due to COVID-19 in week t

$b(t) =$ SARS patients in ICU beds in week t

$c(t) =$ New confirmed COVID-19 patients in regular hospital beds in week t

$d(t) =$ New confirmed COVID-19 patients in ICU beds in week t

$e(t) =$ Active cases in week t

$f(t) =$ Recovered people during the seven weeks prior to t

$g(t) =$ Deaths due to COVID-19 in week t

$h(t) =$ COVID-19 patients in ICU beds in week t

$i(t) =$ Free ICU beds in week t

$B_1^{R_j}(t) = \frac{a(t)}{1+a(t-1)}$; $B_2^{M_{k(j)}}(t) = \frac{b(t)}{1+b(t-1)}$; $B_3^{M_{k(j)}}(t) = \frac{c(t)}{1+c(t-1)}$; $B_4^{M_{k(j)}}(t) = \frac{d(t)}{1+d(t-1)}$; $B_5^{R_j}(t) = \frac{e(t)}{1+f(t)}$;
$B_6^{R_j}(t) = \frac{a(t)\cdot 100{,}000}{P_j}$; $B_7^{R_j}(t) = \frac{g(t)\cdot h(t)}{h(t-1)}$; $B_8^{M_{k(j)}}(t) = \frac{i(t)}{h(t)}$; $B_9^{M_{k(j)}}(t) = \frac{i(t)}{i(t-1)}$; $B_{10}^{S}(t) = \frac{i(t)}{h(t)}$; $B_{11}^{S}(t) = \frac{i(t)}{i(t-1)}$

Each parameter is associated with a 'flag' according to the following ranges:

$$\beta_1^{R_j}(t) = \begin{cases} 0, \text{ if } B_1^{R_j}(t) < 1.05 \\ 1, \text{ if } 1.05 \leq B_1^{R_j}(t) < 1.2 \\ 2, \text{ if } 1.2 \leq B_1^{R_j}(t) < 1.5 \\ 3, \text{ if } 1.5 \leq B_1^{R_j}(t) \end{cases} \qquad \beta_2^{R_j}(t) = \begin{cases} 0, \text{ if } B_2^{M_{k(j)}}(t) < 1.05 \\ 1, \text{ if } 1.05 \leq B_2^{M_{k(j)}}(t) < 1.3 \\ 2, \text{ if } 1.3 \leq B_2^{M_{k(j)}}(t) < 1.5 \\ 3, \text{ if } 1.5 \leq B_2^{M_{k(j)}}(t) \end{cases}$$

$$\beta_3^{R_j}(t) = \begin{cases} 0, \text{ if } B_3^{M_{k(j)}}(t) < 1.05 \\ 1, \text{ if } 1.05 \leq B_3^{M_{k(j)}}(t) < 1.2 \\ 2, \text{ if } 1.2 \leq B_3^{M_{k(j)}}(t) < 1.5 \\ 3, \text{ if } 1.5 \leq B_3^{M_{k(j)}}(t) \end{cases} \qquad \beta_4^{R_j}(t) = \begin{cases} 0, \text{ if } B_4^{M_{k(j)}}(t) < 1.05 \\ 1, \text{ if } 1.05 \leq B_4^{M_{k(j)}}(t) < 1.1 \\ 2, \text{ if } 1.1 \leq B_4^{M_{k(j)}}(t) < 1.25 \\ 3, \text{ if } 1.25 \leq B_4^{M_{k(j)}}(t) \end{cases}$$

$$\beta_5^{R_j}(t)=\begin{cases}0, \text{ if } B_5^{R_j}(t)<0.25\\ 1, \text{ if } 0.25\le B_5^{R_j}(t)<0.5\\ 2, \text{ if } 0.5\le B_5^{R_j}(t)<0.75\\ 3, \text{ if } 0.75\le B_5^{R_j}(t)\end{cases} \qquad \beta_6^{R_j}(t)=\begin{cases}0, \text{ if } B_6^{R_j}(t)<1.5\\ 1, \text{ if } 1.5\le B_6^{R_j}(t)<3\\ 2, \text{ if } 3\le B_6^{R_j}(t)<5\\ 3, \text{ if } 5\le B_6^{R_j}(t)\end{cases}$$

$$\beta_7^{R_j}(t)=\begin{cases}0, \text{ if } B_7^{R_j}(t)<0.25\\ 1, \text{ if } 0.25\le B_7^{R_j}(t)<0.6\\ 2, \text{ if } 0.6\le B_7^{R_j}(t)<1\\ 3, \text{ if } 1\le B_7^{R_j}(t)\end{cases} \qquad \beta_8^{R_j}(t)=\begin{cases}0, \text{ if } 4<B_8^{M_{k(j)}}(t)\\ 1, \text{ if } 2.35<B_8^{M_{k(j)}}(t)\le 4\\ 2, \text{ if } 1.5<B_8^{M_{k(j)}}(t)\le 2.35\\ 3, \text{ if } B_8^{M_{k(j)}}(t)\le 1.5\end{cases}$$

$$\beta_9^{R_j}(t)=\begin{cases}0, \text{ if } 4<B_9^{M_{k(j)}}(t)\\ 1, \text{ if } 2.35<B_9^{M_{k(j)}}(t)\le 4\\ 2, \text{ if } 1.5<B_9^{M_{k(j)}}(t)\le 2.35\\ 3, \text{ if } B_9^{M_{k(j)}}(t)\le 1.5\end{cases} \qquad \beta_{10}^{R_j}(t)=\begin{cases}0, \text{ if } 1.001<B_{10}^{S}(t)\\ 1, \text{ if } 0.8<B_{10}^{S}(t)\le 1.001\\ 2, \text{ if } 0.7<B_{10}^{S}(t)\le 0.8\\ 3, \text{ if } B_{10}^{S}(t)\le 0.7\end{cases}$$

$$\beta_{11}^{R_j}(t)=\begin{cases}0, \text{ if } 1.001<B_{11}^{S}(t)\\ 1, \text{ if } 0.95<B_{11}^{S}(t)\le 1.001\\ 2, \text{ if } 0.8<B_{11}^{S}(t)\le 0.95\\ 3, \text{ if } B_{11}^{S}(t)\le 0.8\end{cases}$$

Fix the weights $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.375, \alpha_5=1, \alpha_6=\alpha_7=\alpha_8=\alpha_9=\alpha_{10}=\alpha_{11}=1.25$. The flag assigned to cities in region j in week t is

$$B^j(t)=\left\lfloor \sum_{n=1}^{11}\alpha_n\cdot\beta_n^{R_j}(t)\right\rfloor.$$

A.2 Our formula

Our formula is computed using the parameters of the government formula that have been obtained for cities.

$C_j=$ Counts the total of municipalities in the cluster $j\in\{1,2,\ldots,20\}$

$B_1^{C_j}(t)=\frac{e(t-1)}{e(t-2)}$; $B_2^{C_j}(t)=\frac{d(t)}{1+d(t-1)}$; $B_3^{C_j}(t)=\frac{e(t)}{1+f(t)}$; $B_4^{C_j}(t)=\frac{g(t)\cdot h(t)}{h(t-1)}$; $B_5^{C_j}(t)=\frac{e(t)\cdot 100{,}000}{P_j}$; $B_6^{C_j}(t)=\frac{i(t)}{h(t)}$; $B_7^{C_j}(t)=\frac{i(t)}{i(t-1)}$; $B_8^{S}(t)=\frac{i(t)}{h(t)}$; $B_9^{S}(t)=\frac{i(t)}{i(t-1)}$

Each parameter is associated with a 'flag' according to the following ranges:

$$\beta_1^{C_j}(t)=\begin{cases}0, \text{ if } B_1^{C_j}(t)<1.05\\ 1, \text{ if } 1.05\le B_1^{C_j}(t)<1.2\\ 2, \text{ if } 1.2\le B_1^{C_j}(t)<1.5\\ 3, \text{ if } 1.5\le B_1^{C_j}(t)\end{cases} \qquad \beta_2^{C_j}(t)=\begin{cases}0, \text{ if } B_2^{C_j}(t)<1.05\\ 1, \text{ if } 1.05\le B_2^{C_j}(t)<1.1\\ 2, \text{ if } 1.1\le B_2^{C_j}(t)<1.25\\ 3, \text{ if } 1.25\le B_2^{C_j}(t)\end{cases}$$

$$\beta_3^{C_j}(t) = \begin{cases} 0, \text{ if } B_3^{C_j}(t) < 0.25 \\ 1, \text{ if } 0.25 \leq B_3^{C_j}(t) < 0.5 \\ 2, \text{ if } 0.5 \leq B_3^{C_j}(t) < 0.75 \\ 3, \text{ if } 0.75 \leq B_3^{C_j}(t) \end{cases} \qquad \beta_4^{C_j}(t) = \begin{cases} 0, \text{ if } B_4^{C_j}(t) < 0.25 \\ 1, \text{ if } 0.25 \leq B_4^{C_j}(t) < 0.6 \\ 2, \text{ if } 0.6 \leq B_4^{C_j}(t) < 1 \\ 3, \text{ if } 1 \leq B_4^{C_j}(t) \end{cases}$$

$$\beta_5^{C_j}(t) = \begin{cases} 0, \text{ if } B_5^{C_j}(t) < 30 \\ 1, \text{ if } 30 \leq B_5^{C_j}(t) < 90 \\ 2, \text{ if } 90 \leq B_5^{C_j}(t) < 270 \\ 3, \text{ if } 270 \leq B_5^{C_j}(t) \end{cases} \qquad \beta_6^{C_j}(t) = \begin{cases} 0, \text{ if } 4 < B_6^{C_j}(t) \\ 1, \text{ if } 2.35 < B_6^{C_j}(t) \leq 4 \\ 2, \text{ if } 1.5 < B_6^{C_j}(t) \leq 2.35 \\ 3, \text{ if } B_6^{C_j}(t) \leq 1.5 \end{cases}$$

$$\beta_7^{C_j}(t) = \begin{cases} 0, \text{ if } 1 < B_7^{C_j}(t) \\ 1, \text{ if } 0.8 < B_7^{C_j}(t) \leq 1 \\ 2, \text{ if } 0.7 < B_7^{C_j}(t) \leq 0.8 \\ 3, \text{ if } B_7^{C_j}(t) \leq 0.7 \end{cases} \qquad \beta_8^{C_j}(t) = \begin{cases} 0, \text{ if } 4 < B_8^{S}(t) \\ 1, \text{ if } 2.35 < B_8^{S}(t) \leq 4 \\ 2, \text{ if } 1.5 < B_8^{S}(t) \leq 2.35 \\ 3, \text{ if } B_8^{S}(t) \leq 1.5 \end{cases}$$

$$\beta_9^{C_j}(t) = \begin{cases} 0, \text{ if } 1.001 < B_9^{S}(t) \\ 1, \text{ if } 0.95 < B_9^{S}(t) \leq 1.001 \\ 2, \text{ if } 0.8 < B_9^{S}(t) \leq 0.95 \\ 3, \text{ if } B_9^{S}(t) \leq 0.8 \end{cases}$$

Fix the weights $\alpha_1 = \alpha_2 = 0.75, \alpha_3 = 1, \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = 1.25$, the flag assigned to the cities in cluster j in week t is

$$B^j(t) = \left\lfloor \sum_{n=1}^{9} \alpha_n \cdot \beta_n^{C_j}(t) \right\rfloor$$