Co-clustering analysis of Covid-19 in the U.S.

Abstract

In response to the global coronavirus pandemic, we study co-clustering of multivariate time-series data as a way to simultaneously cluster both geographical regions and time periods after the outbreak of the pandemic. The resulting blocks of clusters, identified with a latent block model component, are integrated with an extended SIR model that takes both geographical clusters as well as clusters of time periods into account. We consider US data (or global or both?) and show how the novel latent block SIR model produce better prediction accuracy of the epidemic, and gives further understanding of how the virus is spreading in different geographical regions during different periods of time.

1 Introduction

Following the advances of modern technology, it is nowadays not difficult to collect high frequency data. Smart phones, smart watches and similar devices for example makes it possible to gather a vast amount of data on each individual over time. With such data collected, it is of interest to summarize the data by homogeneous subgroups of the data matrix to, for example, increase individual prediction accuracy. Methods for clustering and classification of functional data has rapidly increased over time, and according to Jacques and Preda 2014a, there are four groups of clustering techniques for functional data: 1) the raw data methods that cluster the functions directly on the curves' finite set of point, 2) the filtering methods that first smooths the curves into basis functions and then clusters the basis expansion coefficients, 3) the adaptive methods that simultaneously cluster and express the functions in a finite dimensional space, and 4) the distance-based methods where classical algorithmic approaches such as k-means and hierarchical clustering are adapted to functional data.

For all of the above clustering approaches there exist many studies considering univariate functional data $X(t)$. For multivariate functional data however, the number of studies are much fewer. Both Singhal and Seborg 2005 and Ieva et al. 2013 propose k-means clustering extended for multivariate functional data. This study however takes on a model-based approach to clustering using mixture models.

The aim of this paper is to study the coronavirus pandemic, both on a global level and on a US level. A joint analysis of the number of cases and deaths is conducted in order to identify spatio-temporal clusters that captures the trend of the pandemic. We thus aim to cluster both the rows (the states) and the columns (the weeks) in the data matrix. In the statistical literature, this is known as co-clustering Bouveyron et al. 2018. The analysis and theory presented assumes data in the form of multivariate functions, meaning that the data in itself are infinite dimensional. Some challenging problems for both theory and computation follows, which will be addressed as well. In the univariate

case, each individual (i.e., country for this analysis) is described by a single function (stochastic process) $X(t) \in \mathbb{R}$, $\forall t[0,T]$, whereas for the multivariate case $\mathbf{X} = \mathbf{X}(t)_{t\in[0,T]}$ where $\mathbf{X}(t) = (X^1(t), \ldots, X^p(t))^\top$.

In a second stage of the analysis, the resulting blocks from the co-clustering of step one will be plugged into the well-known epidemiological SIR model. With information from the co-clusters, it is possible to make predictions about the pandemic for larger, homogeneous groups of countries. It is furthermore possible to enhance predictions by using the column clusters of time periods for which the pandemic spread in similar ways in the different row (country) clusters. Lastly, an important contribution of this study is the inclusion of dependence between observations in the recovery of the functional form of the data before the co-clustering is performed.

QUESTION: How to formulate a model for wavelet smoothing in the case of peaked data?

2 Functional Data Analysis

We consider measurable, real-valued functions $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that are realisations of a continuous, multivariate stochastic process $\mathbf{X} = \{\mathbf{X}(t),\, t \in [0,T]\} = \{(X^1(t), \ldots, X^p(t))\}_{t\in[0,T]}$ that belongs to the separable Hilbert space $L_2([0,1], \mathbb{R})$ of square integrable functions on $[0,1]$. The sample paths are denoted by $\mathbf{X}_i = (X_i^1, \ldots, X_i^p) \in L_2(I), I = (0,T), 1 \leq i \leq n$.

In many cases, the functions will only be observed at discrete time points on a fixed time grid, $X^j(t_1), \ldots, X^j(t_s), 0 \leq t_1 \leq \ldots \leq t_s \leq 1 \forall 1 \leq i \leq n, 1 \leq j \leq p$. The first task thus becomes to recover the functional form from the discrete observations. With measurement errors often present in real data, smoothing techniques are often employed for this mean, assuming that the errors fluctuate around a smooth trajectory. A common approach is to make the assumption that the functions $X_i^j(t)$ can be decomposed into a finite-dimensional space that is spanned by a set of basis functions. Denoting the basis functions of the jth component as $(\phi_r^j(t))_{1\leq r\leq R_j}$ and R_j the number of basis functions,

$$X_i^j(t) = \sum_{r=1}^{R_j} c_{ir}^j(X_i^j)\phi_r^j(t) \tag{1}$$

Lastly, we gather the c_{ir}^j coefficients and the ϕ_r^j basis functions in the matrices $\mathbf{C}$ and $\boldsymbol{\phi}$, where

$$\mathbf{C} = \begin{pmatrix} c_{11}^1 & \ldots & c_{1R_1}^1 & c_{11}^2 & \ldots & c_{1R_2}^2 & c_{11}^p & \ldots & c_{1R_p}^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{n1}^1 & \ldots & c_{nR_1}^1 & c_{n1}^2 & \ldots & c_{nR_2}^2 & c_{n1}^p & \ldots & c_{nR_p}^p \end{pmatrix}$$

and

$$\boldsymbol{\phi}(\mathbf{t}) = \begin{pmatrix} \phi_1^1(t) & \ldots & \phi_{R_1}^1(t) & 0 & \ldots & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \phi_1^2(t) & \ldots & \phi_{R_2}^2(t) & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & \phi_1^p(t) & \ldots & \phi_{R_p}^p(t) \end{pmatrix}$$

where the rth column equals the values of the rth basis function at time points $\mathbf{t}^\top = [t_1, t_2, \ldots, t_s]$. With matrix notation, we can now express Equation (1) as

$$\mathbf{X}(t) = \mathbf{C}\boldsymbol{\phi}^{\top}(t)$$

Estimating the matrix $\mathbf{C}$ is often done using least squares JO Ramsay and B. Silverman 2002. Which basis to use, and how many basis functions to include is currently an open problem. It has however been recommended to use Fourier basis for data with a repetitive patterns, and a B-spline basis for most other cases Schmutz et al. 2020. Regarding the number of basis functions to include has to be decided by the user as there are no clear rules on how to select it Jacques and Preda 2014a.

3 Functional Co-Clustering

3.1 Dependent Functional Data Analysis

QUESTION: How to incorporate dependency between observations?

In functional data analysis, the second order structure of the functions are of central importance Panaretos, Tavakoli, et al. 2013. For independent and identically distributed functional data, such structure is completely described by the covariance operator Grenander 1981. More recently, studies considering dependent functional data have emerged.

3.2 The Functional Latent Block Model

The functional latent block model (FLBM, Bouveyron et al. 2018) has been implemented to co-cluster the cases/deaths and the weeks. On a general level, the FLBM assumes the data matrix $\mathbf{x} = (\mathbf{x}_{ij})$, $i = 1, ..., n$, $j = 1, ..., p$, where each entry is a multivariate curve $\mathbf{x}_{ij} = (x_{ij}^1(t), \ldots, x_{ij}^S(t))$, $t \in [0, T]$ and where s is the component of the multivariate curves. For the analyzed Covid-19 data, there are thus $n = 50$ states, $p = 12$ weeks and s represents either cases or deaths. Since the functional form of the curves $x_{ij}^s(t)$ is unknown, it is also assumed that the curves belong to a finite-dimensional space spanned by basis functions $\{\phi_r^s\}_{r=1,\ldots,R_s}$ and that they can be expressed as linear combinations of those basis functions:

$$x_{ij}^s(t) = \sum_{r=1}^{R_s} c_{ijr}^s \phi_r^s(t), \tag{2}$$

where $(c_{ijr}) = \mathbf{c}_{ij}$ denotes the basis expansion coefficients of each curve. The basis functions can for example be Fourier or spline bases, but it is yet unknown how to automatically select basis function and the number of basis functions Jacques and Preda 2014b. The coefficients c_{ijr}^s are often estimated using least squares smoothing James Ramsay and Bernard W. Silverman 2005.

The most common model for co-clustering is the latent block model (LBM, Govaert and Nadif 2013). Let $\mathbf{z} = (z_{ik})$, $j = 1, ..., p$, $k = 1, ..., K$ denote a binary random variable such that $z_{ik} = 1$ if individual i belongs to row cluster k, and $\mathbf{w} = (w_{jl})$, $j = 1, ..., p$, $l = 1, ..., L$ denote a binary random variable such that $w_{jl} = 1$ if feature j belongs to column cluster l. Co-clustering will yield subgroups, called blocks, such that $z_{ik}w_{jl} = 1$. It is assumed that $\mathbf{z}$ and $\mathbf{w}$ are independent from each other and that the random variables $\mathbf{x}$ are independent conditional on $\mathbf{z}$ and $\mathbf{w}$.

From the independence assumption regarding $\mathbf{z}$ and $\mathbf{w}$, it follows that

$$p(c;\theta) = \sum_{z\in\mathbf{Z}} \sum_{w\in\mathbf{W}} p(z;\theta)p(w;\theta)p(c|z,w;\theta) \tag{3}$$

where $c = (c_{ij})_{ij}$. Now let $\alpha_k \in [0,1]$ and $\beta_l \in [0,1]$ denote the respective row and column mixing proportions such that they both sum to 1 and $p(z;\theta) = \prod_{ik} \alpha_k^{z_{ik}}$ and $p(w;\theta) = \prod_{ik} \beta_l^{w_{jl}}$. If it further is assumed that the c_{ij} basis expansion coefficients are independent and generated by a distribution that is block-specific, $p(c|z,w;\theta) = \prod_{ijkl} p(c_{ij};\theta_{kl})^{z_{ik}w_{jl}}$, the distribution in (3) can be expressed as

$$p(c;\theta) = \sum_{z\in\mathbf{Z}} \sum_{w\in\mathbf{W}} \prod_{ik} \alpha_k^{z_{ik}} \prod_{ik} \beta_l^{w_{jl}} \prod_{ijkl} p(c_{ij};\theta_{kl})^{z_{ik}w_{jl}}. \tag{4}$$

4 Estimation

Using the LBM to cluster the rows and columns implies a data structure with missing values on two variables, z and w. It therefore follows that the commonly used EM algorithm does not work since there will be too many terms to compute in the E-step. This study therefore implements the estimation technique suggested in Bouveyron et al. 2018 by using a stochastic version of the EM-algorithm and embedding a Gibbs sampler that generates z and w. In this way it is possible to circumvent having to compute their joint probability distribution. Since the algorithm needs initial values for the column partitions and the parameters, k-means clustering is used to initialize the row and column partitions. From these partitions it is possible to deduce the initial parameter values. Lastly, the number of row and column clusters are chosen by maximizing the Integrated Classification Likelihood (ICL) value, see Bouveyron et al. 2018 for the details.

5 Data

In Table 1, the data is summarized, together with information about when the first Covid-19 case was registered at each state. This date differs a lot between states; for example, in Alabama, the first case was reported at the 13th of March, wheras the first case in California was reported at the 25th of January. The first case in the country was reported in Washington at the 21st of January, however most states had their first reported case in early March. For all states, there are no "missing values" after the date of a first registered case, meaning there are confirmed cases each day for every state starting from the first day of case registration.

In Table 1, there is also a column with the number of deaths per case. This information is also contained in Figure 1, where the number of deaths in relation to state population size is plotted against the number of deaths in relation to the number of cases. The mortality of the virus for the different states appears a bit different compared to only looking at death numbers relative to state population size. For states with a low fraction of deaths relative to state population size, the mortality for confirmed cases is much higher compared to the states with a relatively larger fraction of deaths.

Figure 2 and 3 show the trajectories of deaths and cases over the whole period, starting from the first registered case in January. The trajectories show each state, with color indicating which row cluster it belongs to. Figure 4 shows the five row clusters as different panels.

In Figure 5, the total number of cases are given for each state. As is apparent, the state of New York has almost 4 000 000 infected people, compared to the state average of 222 115 people.

In Figure 6, the total number of deaths are displayed for each state. The state average equals 8410.3 deaths.

	State	Abbr.	Cases	Deaths	Deaths of cases	Fraction cases	Fraction deaths	First case
1	Alabama	AL	559482	20475	0.04	0.11	0.00	2020-03-13
2	Alaska	AK	22826	447	0.02	0.03	0.00	2020-03-12
3	Arizona	AZ	616257	27696	0.04	0.08	0.00	2020-01-26
4	Arkansas	AR	230964	4591	0.02	0.08	0.00	2020-03-11
5	California	CA	3751385	143726	0.04	0.09	0.00	2020-01-25
6	Colorado	CO	1023645	52079	0.05	0.18	0.01	2020-03-05
7	Connecticut	CT	1734813	142974	0.08	0.49	0.04	2020-03-08
8	Delaware	DE	327783	11368	0.03	0.34	0.01	2020-03-11
9	Florida	FL	2275776	87198	0.04	0.11	0.00	2020-03-01
10	Georgia	GA	1723453	73906	0.04	0.16	0.01	2020-03-02
11	Hawaii	HI	37820	851	0.02	0.03	0.00	2020-03-06
12	Idaho	ID	132183	3647	0.03	0.07	0.00	2020-03-13
13	Illinois	IL	4026956	175007	0.04	0.32	0.01	2020-01-24
14	Indiana	IN	1268191	75820	0.06	0.19	0.01	2020-03-06
15	Iowa	IA	587883	14438	0.02	0.19	0.00	2020-03-08
16	Kansas	KS	329552	8648	0.03	0.11	0.00	2020-03-07
17	Kentucky	KY	343811	16175	0.05	0.08	0.00	2020-03-06
18	Louisiana	LA	1838336	118310	0.06	0.40	0.03	2020-03-09
19	Maine	ME	83770	3314	0.04	0.06	0.00	2020-03-12
20	Maryland	MD	1691454	80878	0.05	0.28	0.01	2020-03-05
21	Massachusetts	MA	3898291	235391	0.06	0.57	0.03	2020-02-01
22	Michigan	MI	2591288	226105	0.09	0.26	0.02	2020-03-10
23	Minnesota	MN	615792	28982	0.05	0.11	0.01	2020-03-06
24	Mississippi	MS	507678	22152	0.04	0.17	0.01	2020-03-11
25	Missouri	MO	526853	25426	0.05	0.09	0.00	2020-03-07
26	Montana	MT	28879	845	0.03	0.03	0.00	2020-03-13
27	Nebraska	NE	396122	5279	0.01	0.20	0.00	2020-02-17
28	Nevada	NV	344223	16099	0.05	0.11	0.01	2020-03-05
29	New Hampshire	NH	162265	6877	0.04	0.12	0.01	2020-03-02
30	New Jersey	NJ	7234185	448178	0.06	0.81	0.05	2020-03-04
31	New Mexico	NM	253369	10331	0.04	0.12	0.00	2020-03-11
32	New York	NY	18775939	1376727	0.07	0.97	0.07	2020-03-01
33	North Carolina	NC	840922	28059	0.03	0.08	0.00	2020-03-03
34	North Dakota	ND	81356	1836	0.02	0.11	0.00	2020-03-11
35	Ohio	OH	1255910	69208	0.06	0.11	0.01	2020-03-09
36	Oklahoma	OK	251520	13684	0.05	0.06	0.00	2020-03-06
37	Oregon	OR	176614	6589	0.04	0.04	0.00	2020-02-28
38	Pennsylvania	PA	3072958	181844	0.06	0.24	0.01	2020-03-06
39	Rhode Island	RI	548757	21141	0.04	0.52	0.02	2020-03-01
40	South Carolina	SC	426778	16664	0.04	0.08	0.00	2020-03-06
41	South Dakota	SD	171116	1698	0.01	0.19	0.00	2020-03-10
42	Tennessee	TN	808879	13773	0.02	0.12	0.00	2020-03-05
43	Texas	TX	2181099	57768	0.03	0.08	0.00	2020-02-12
44	Utah	UT	340148	3557	0.01	0.11	0.00	2020-02-25
45	Vermont	VT	54998	2867	0.05	0.09	0.00	2020-03-07
46	Virginia	VA	1294819	42147	0.03	0.15	0.00	2020-03-07
47	Washington	WA	1030357	53549	0.05	0.14	0.01	2020-01-21
48	West Virginia	WV	75760	2701	0.04	0.04	0.00	2020-03-17
49	Wisconsin	WI	570738	21283	0.04	0.10	0.00	2020-02-05
50	Wyoming	WY	37265	433	0.01	0.06	0.00	2020-03-11

Table 1: The number of cases, deaths and fraction of Covid-19 cases per state together with the date of the first reported Covid-19 case.

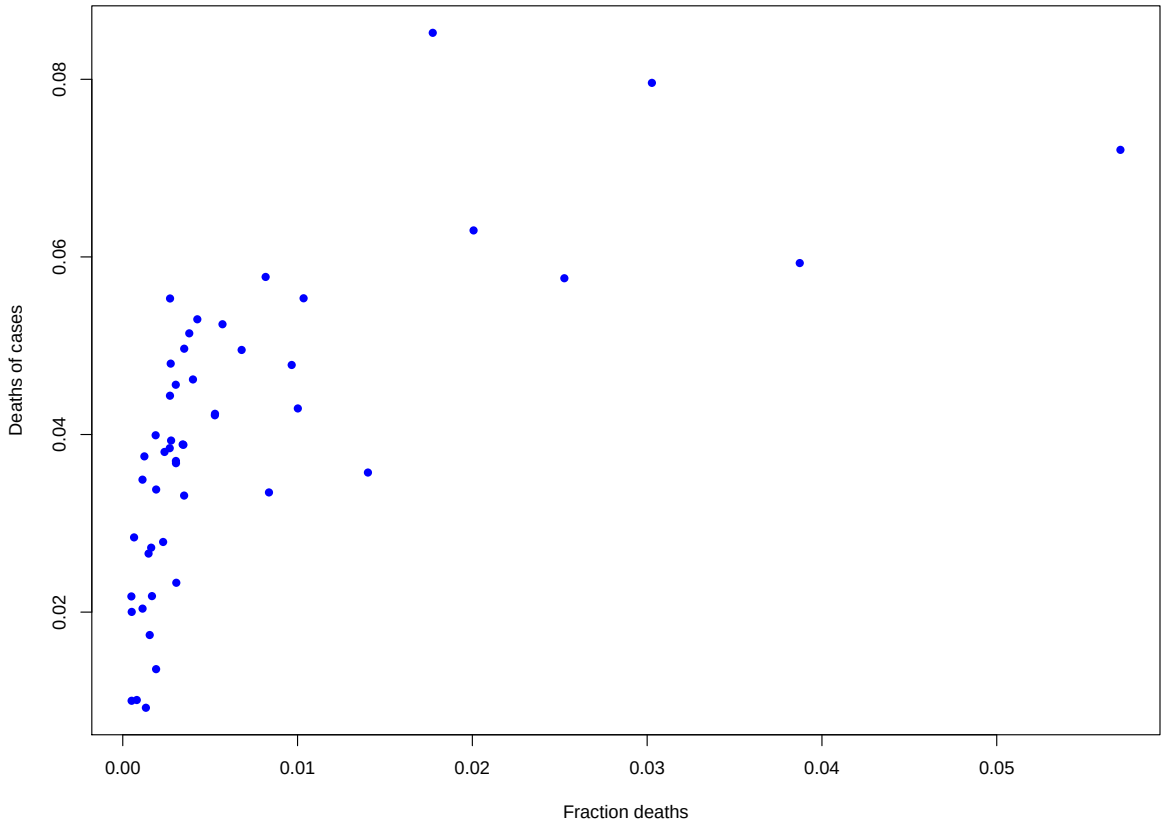

Figure 1: Fraction of deaths plotted against fraction of cases.

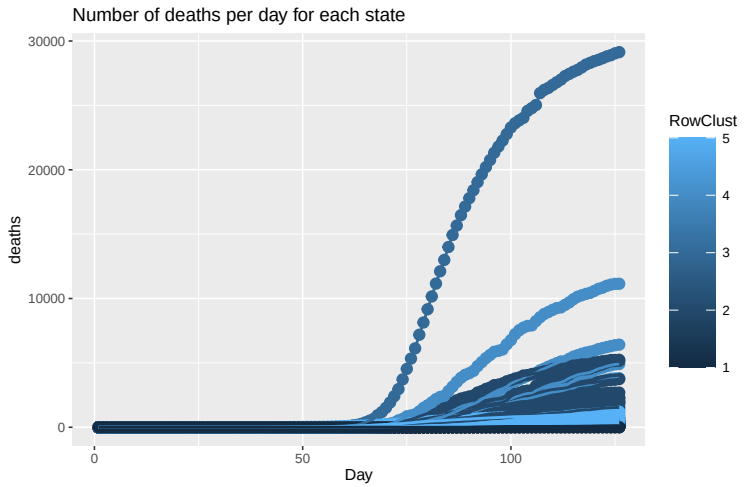

Figure 2: Trajectories of deaths per day for each state.

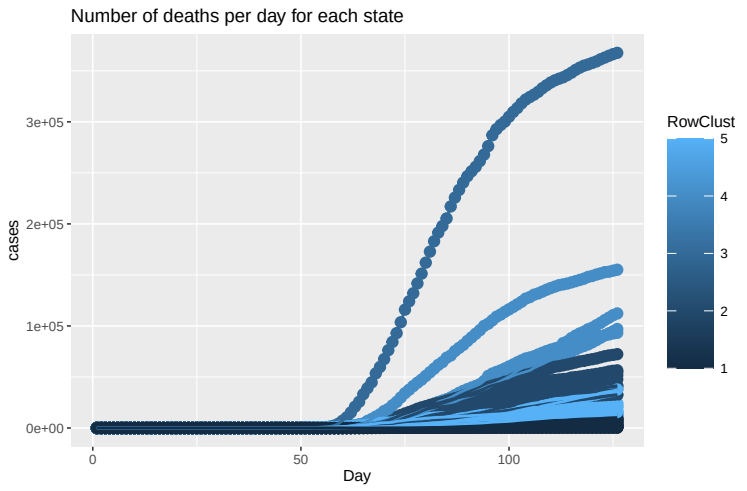

Figure 3: Trajectories of deaths per day for each state.

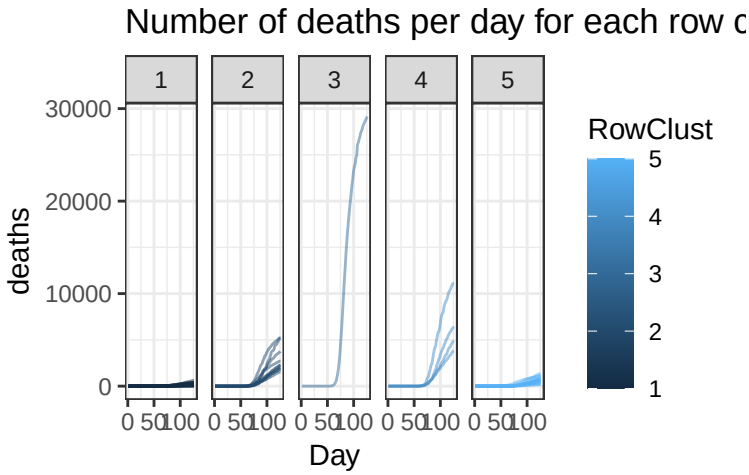

Figure 4: Trajectories of deaths per day for each row cluster.

In Figure 5 and 6 the number of cases and deaths per state are displayed. It is clear that the state of New York and state of New Jersey stand out.

6 Co-Clustering Results

6.1 Univariate co-clustering of Covid-19 cases

In Figure 7 the block mean functions for the variable "Cases" is shown. For the row clusters 1 and 5, there are no period variations. For row cluster three there are on the otherside distinct differences for the different column clusters.

In Figure 8 the column clusters are displayed over the measurement period. There is a clear division between the clusters, with March being almost entirely represented by column cluster 1 and May by column cluster 5. In April, clusters 2 and 3 are representing

half of the month each. There thus seem to be distinct periods in the virus outbreak, where the spread of the virus continuously enters new phases along with time.

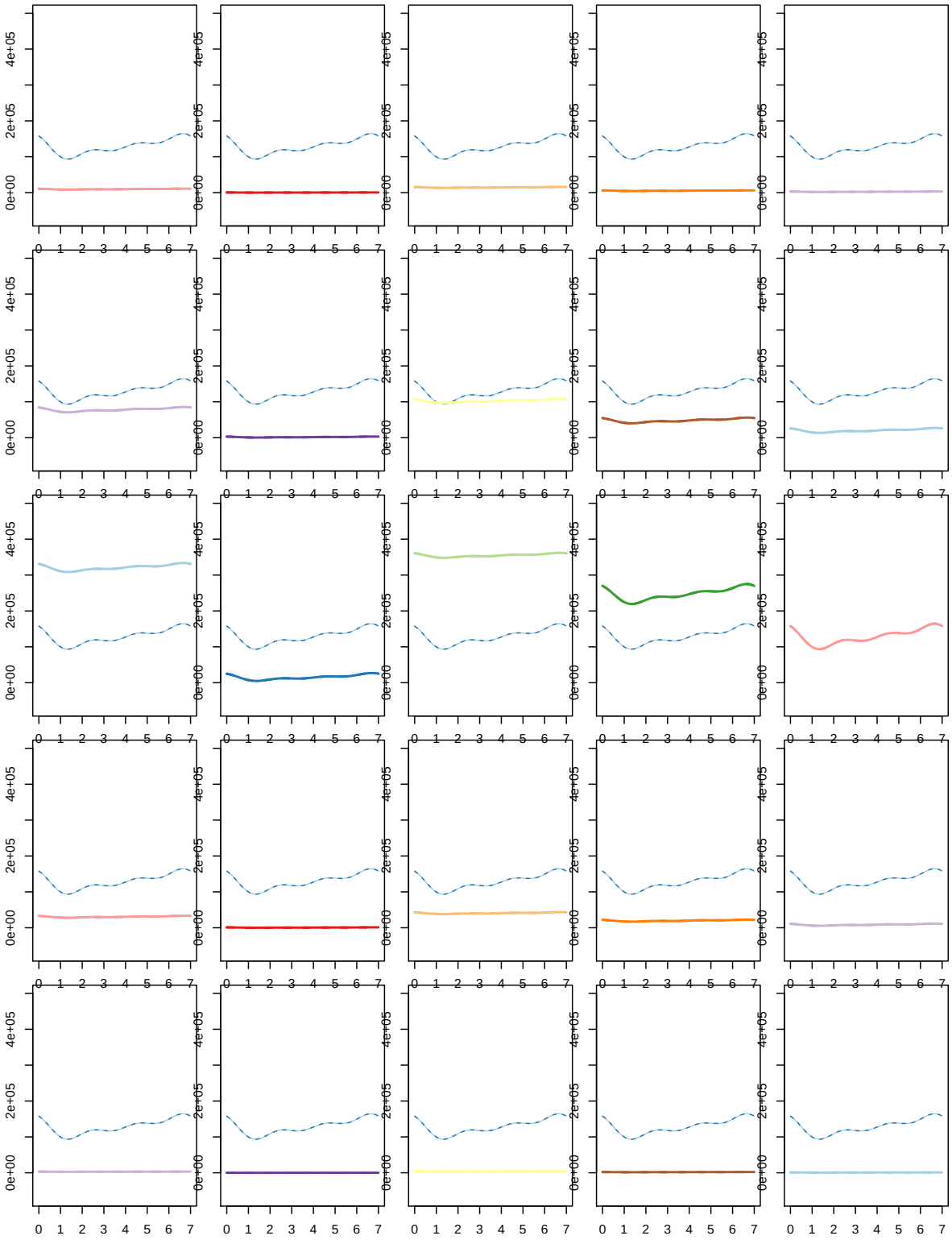

Figure 7: Blocks showing the mean functions for the univariate co-clustering of Covid-19 cases.

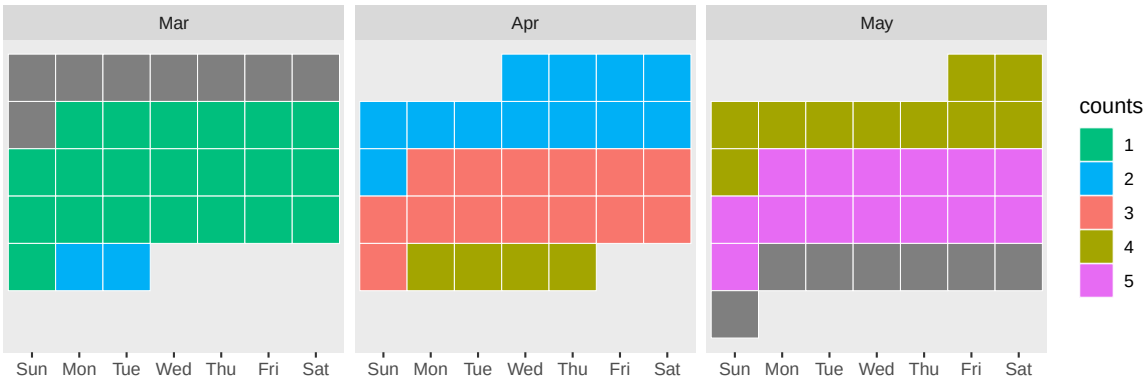

Figure 8: Calendar showing the column clusters for the univariate co-clustering of Covid-19 cases.

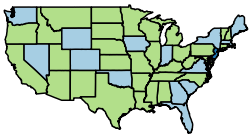

Figure 9: A map of the american states showing the row clusters.

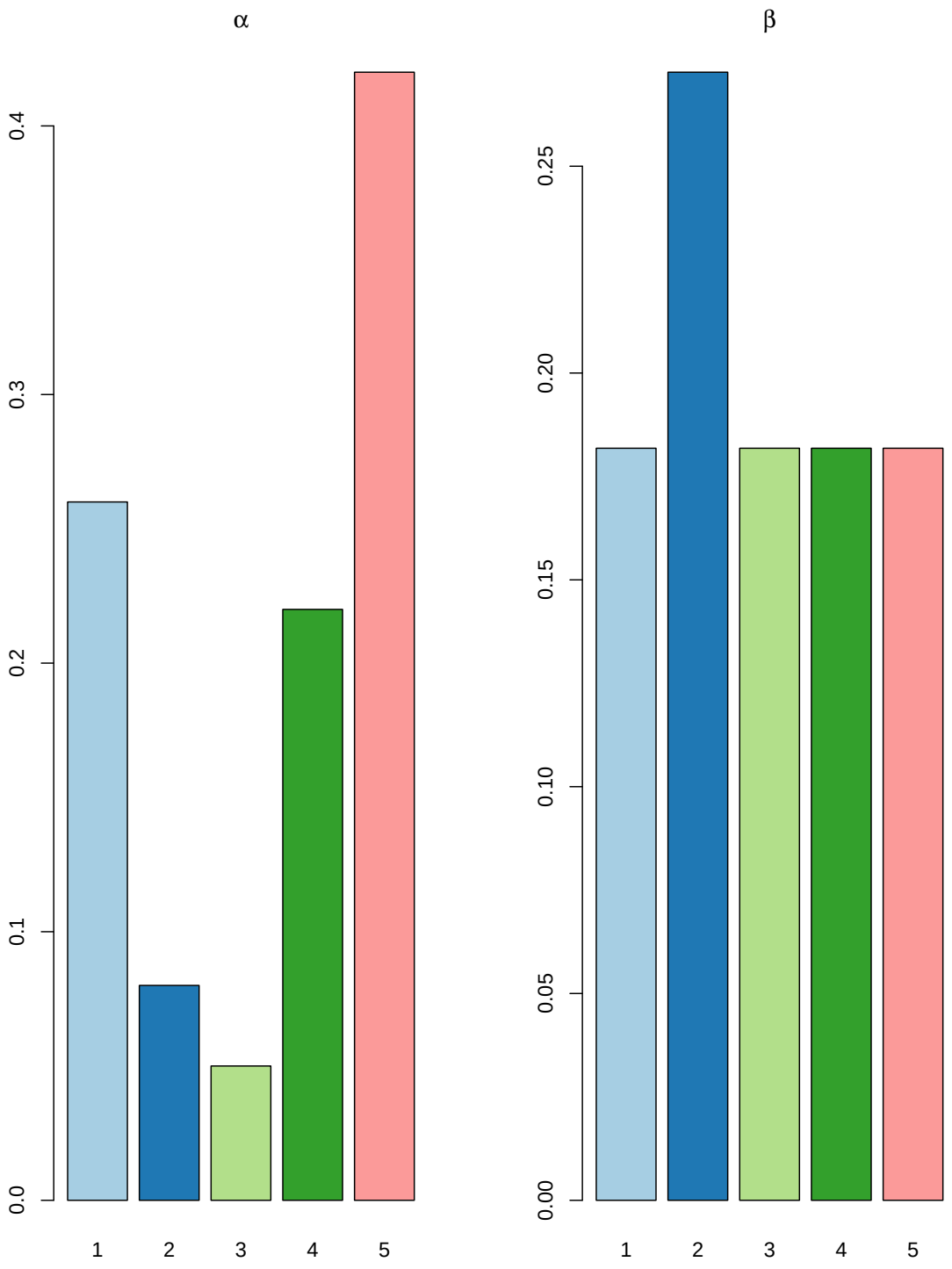

Figure 10: The proportions of states and weeks belonging to the row and column clusters, respectively.

6.2 Univariate co-clustering of Covid-19 deaths

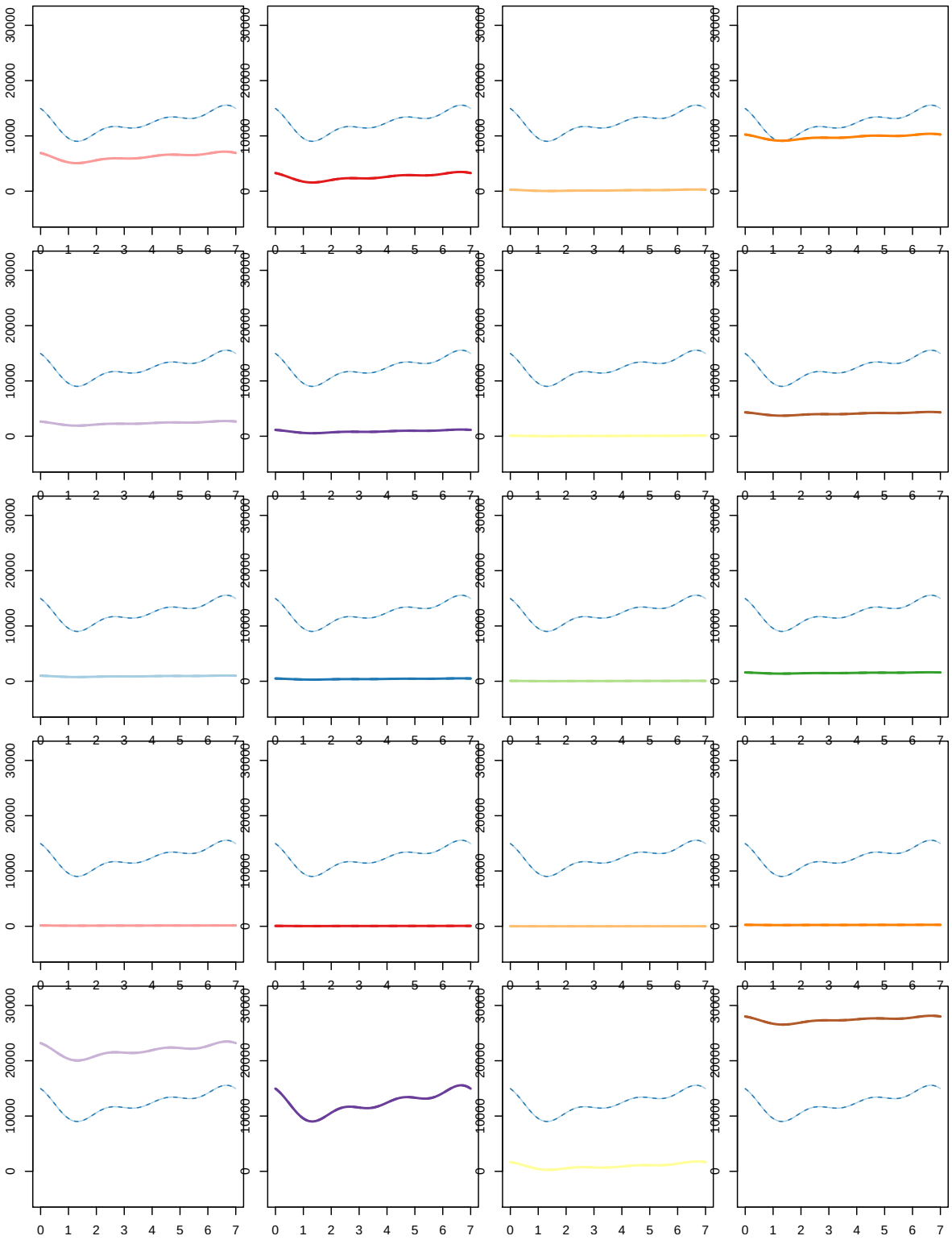

Figure 11: Blocks showing the mean functions for the univariate co-clustering of Covid-19 cases.

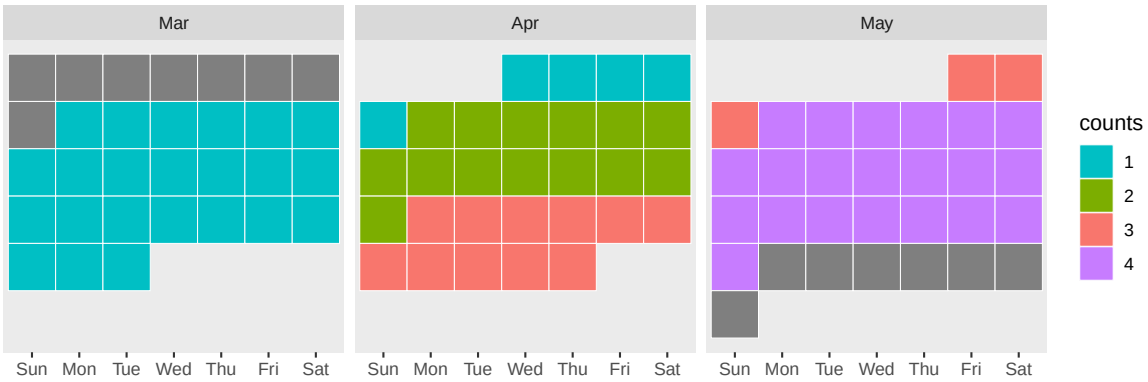

Figure 12: Calendar showing the column clusters for the univariate co-clustering of Covid-19 cases.

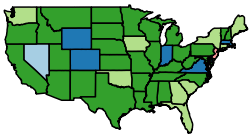

Figure 13: A map of the american states showing the row clusters.

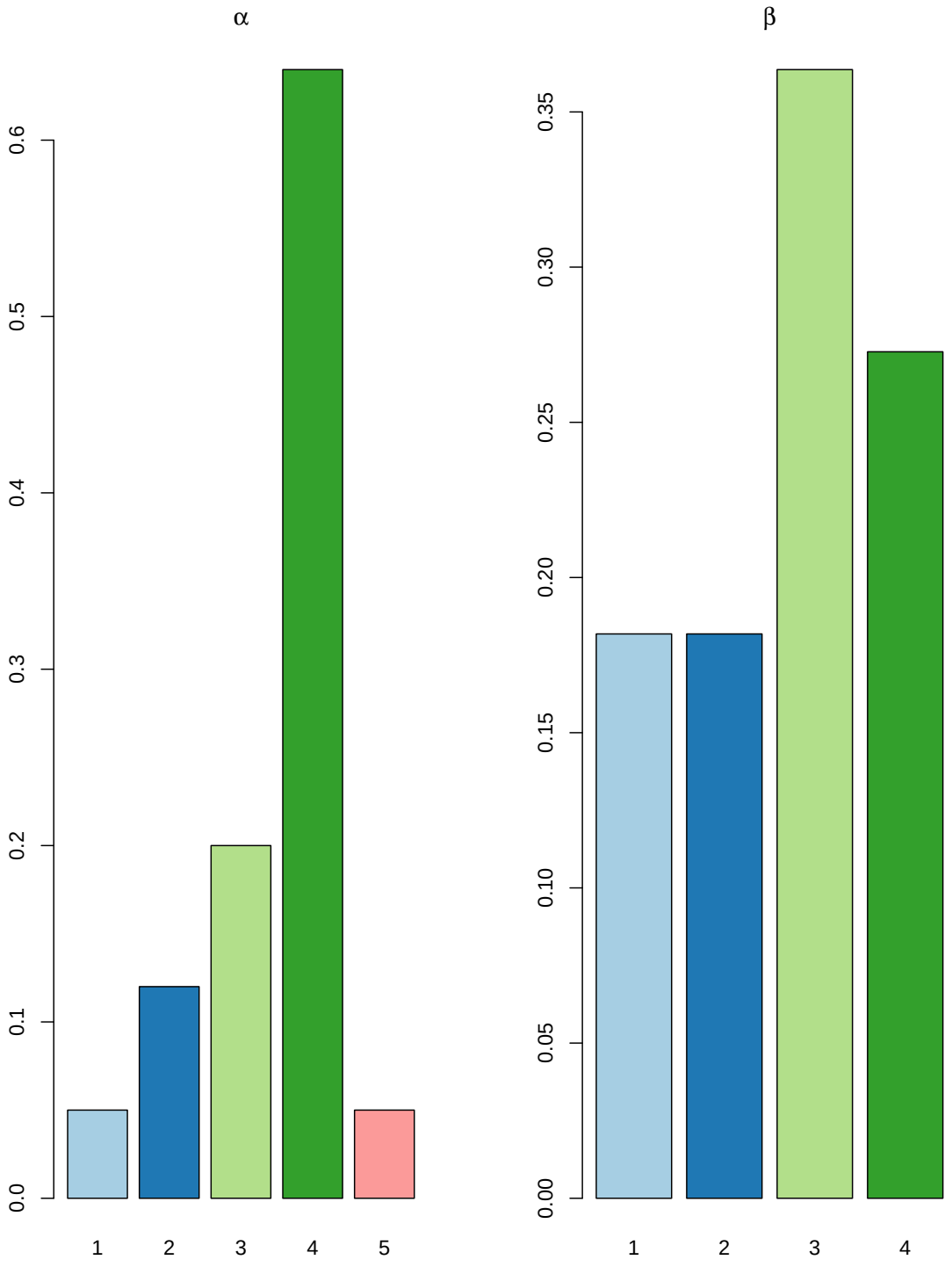

Figure 14: The proportions of states and weeks belonging to the row and column clusters, respectively.

In the next step, a multivariate function considering both deaths and cases is adopted.

6.3 Multivariate co-clustering of Covid-19 cases and deaths

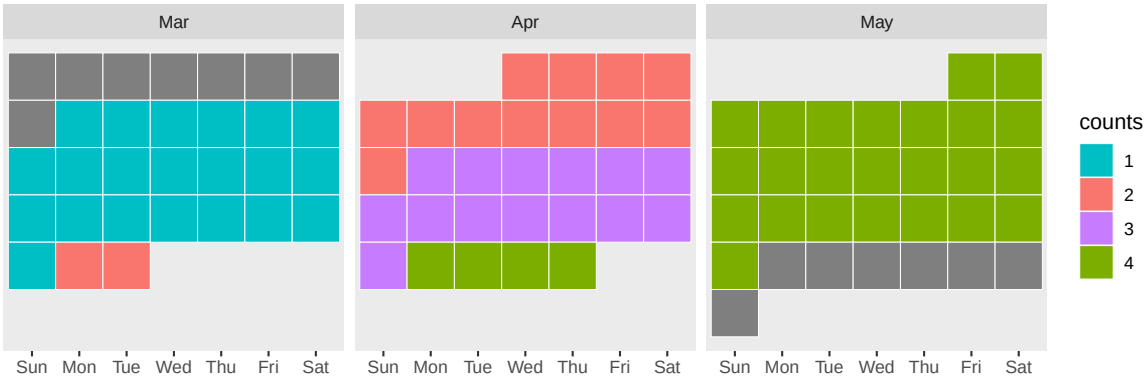

Figure 15: Calendar showing the column clusters for the multivariate co-clustering of Covid-19 cases and deaths.

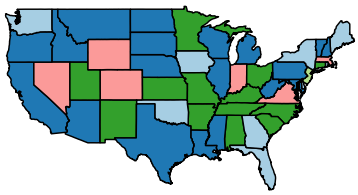

Figure 16: A map of the american states showing the row clusters.

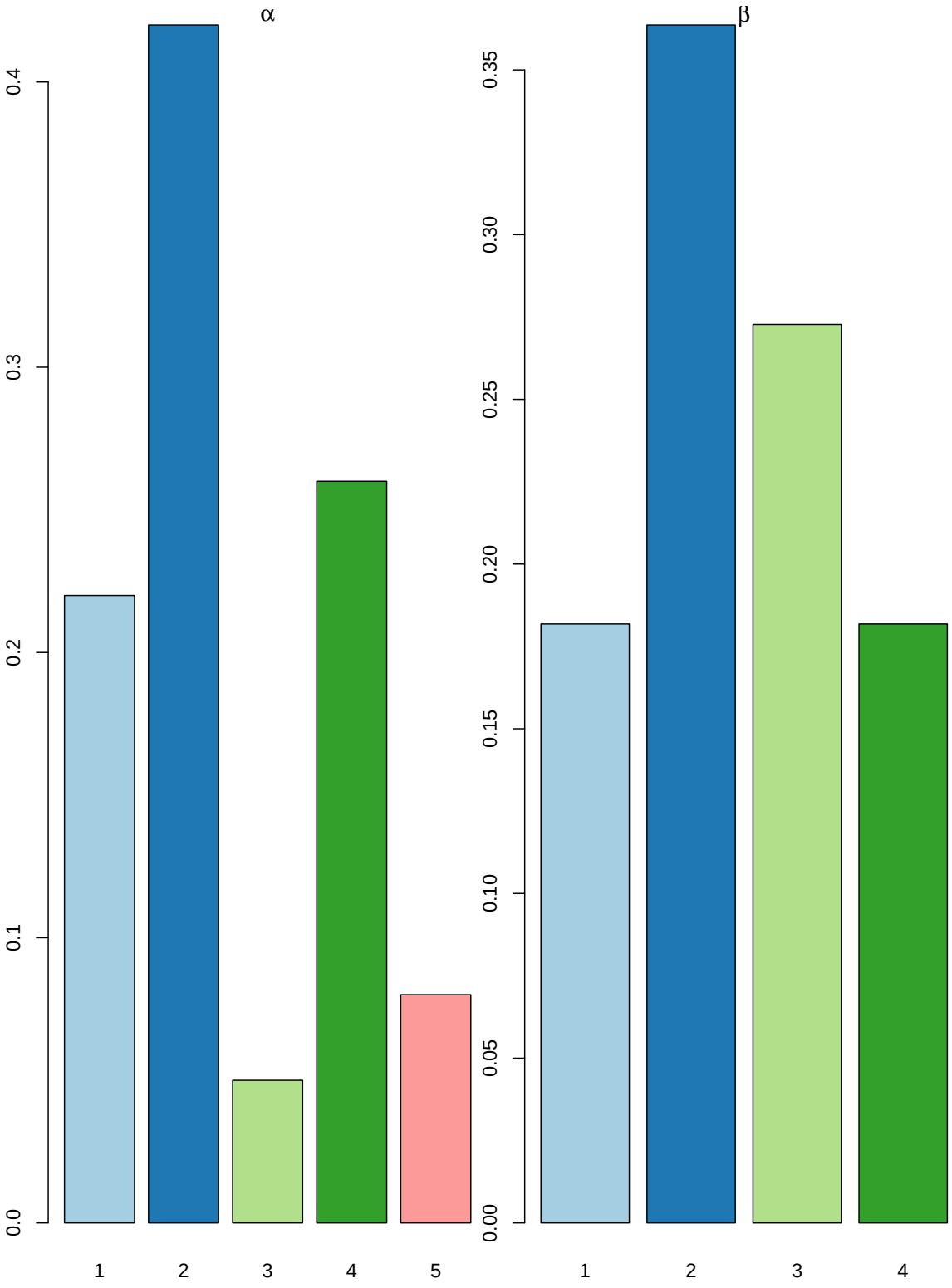

Figure 17: The proportions of states and weeks belonging to the row and column clusters, respectively.

7 Epidemiology Model for Prediction

SIR model.

8 Current problems

Need to fix the block plots, scale the figures and comment them. Need to produce block plot for multivariate case. Need to figure out how the prediction step should look like. How does the SIR model work? What can be done? What has been done in previous studies? Is it possible to incorporate other information to make better predictions? What kind of information is available?

References

Bouveyron, C. et al. (2018). "The Functional Latent Block Model for theCo-Clustering of Electricity Consumption Curves". In: *Journal of the Royal Statistical Society, Series C Applied Statistics* 67, pp. 897–915.

Govaert, G. and M. Nadif (2013). *Co-Clustering: Models, Algorithms and Applications*. Wiley.

Grenander, Ulf (1981). *Abstract inference*. Tech. rep.

Ieva, F. et al. (2013). "Multivariate Functional Clustering for the Morphological Analysis of ECG Curves". In: *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 3.62, pp. 401–418.

Jacques, J. and C. Preda (2014a). "Functional data clustering: a survey". In: *Advances in Data Analysis and Classification* 3.8, pp. 231–255.

— (2014b). "Model-based clustering for multivariate functional data". In: *Computational Statistics and Data Analysis*.

Panaretos, Victor M, Shahin Tavakoli, et al. (2013). "Fourier analysis of stationary time series in function space". In: *The Annals of Statistics* 41.2, pp. 568–603.

Ramsay, James and Bernard W. Silverman (2005). *Functional data analysis*. Springer.

Ramsay, JO and BW Silverman (2002). *Functional Data Analysis-Methods and Case Studies*.

Schmutz, Amandine et al. (2020). "Clustering multivariate functional data in group-specific functional subspaces". In: *Computational Statistics*, pp. 1–31.

Singhal, A. and D. Seborg (2005). "Clustering multivariate time-series data". In: *Journal of Chemometrics* 19, pp. 427–438.

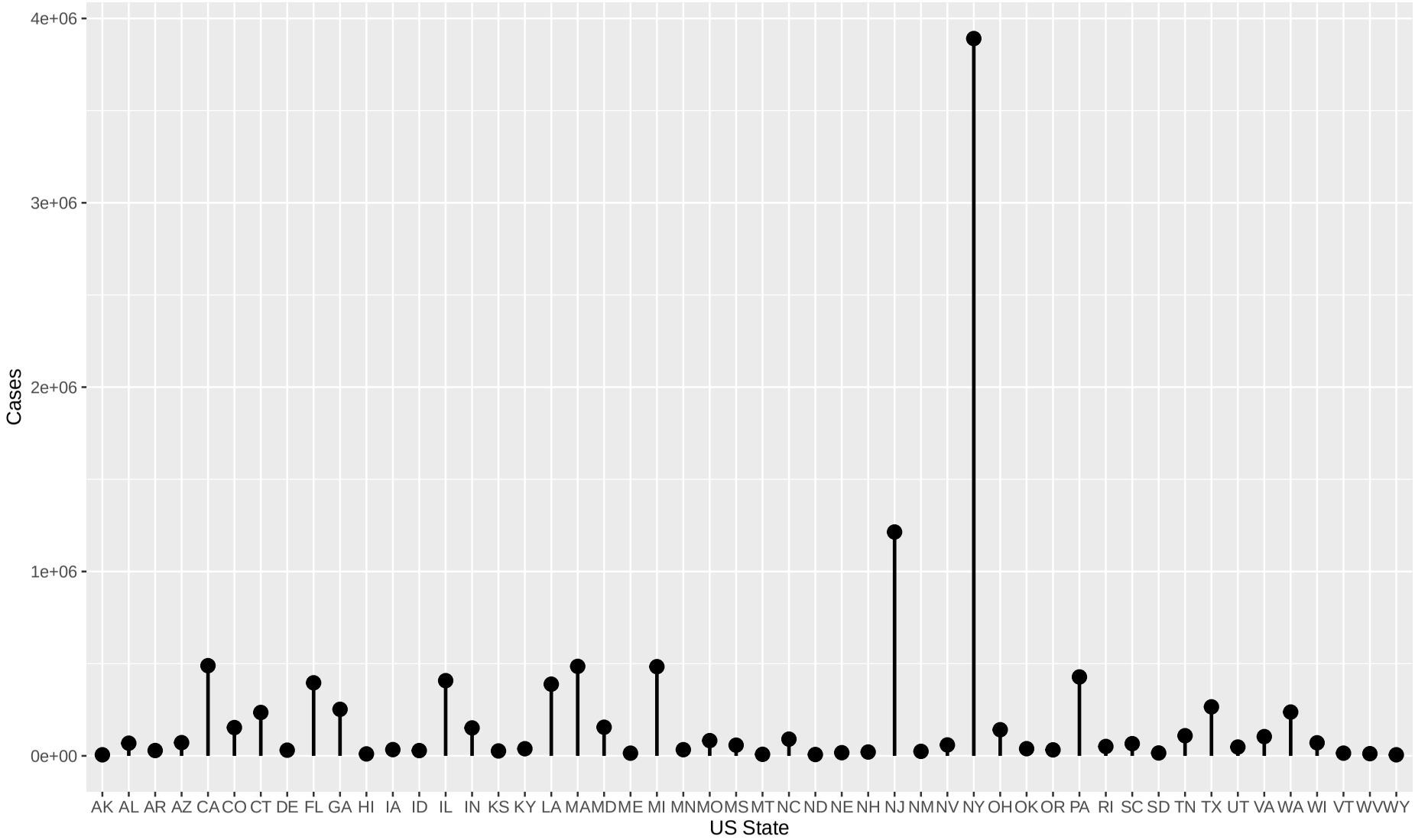

Figure 5: The number of Covid-19 cases per state.

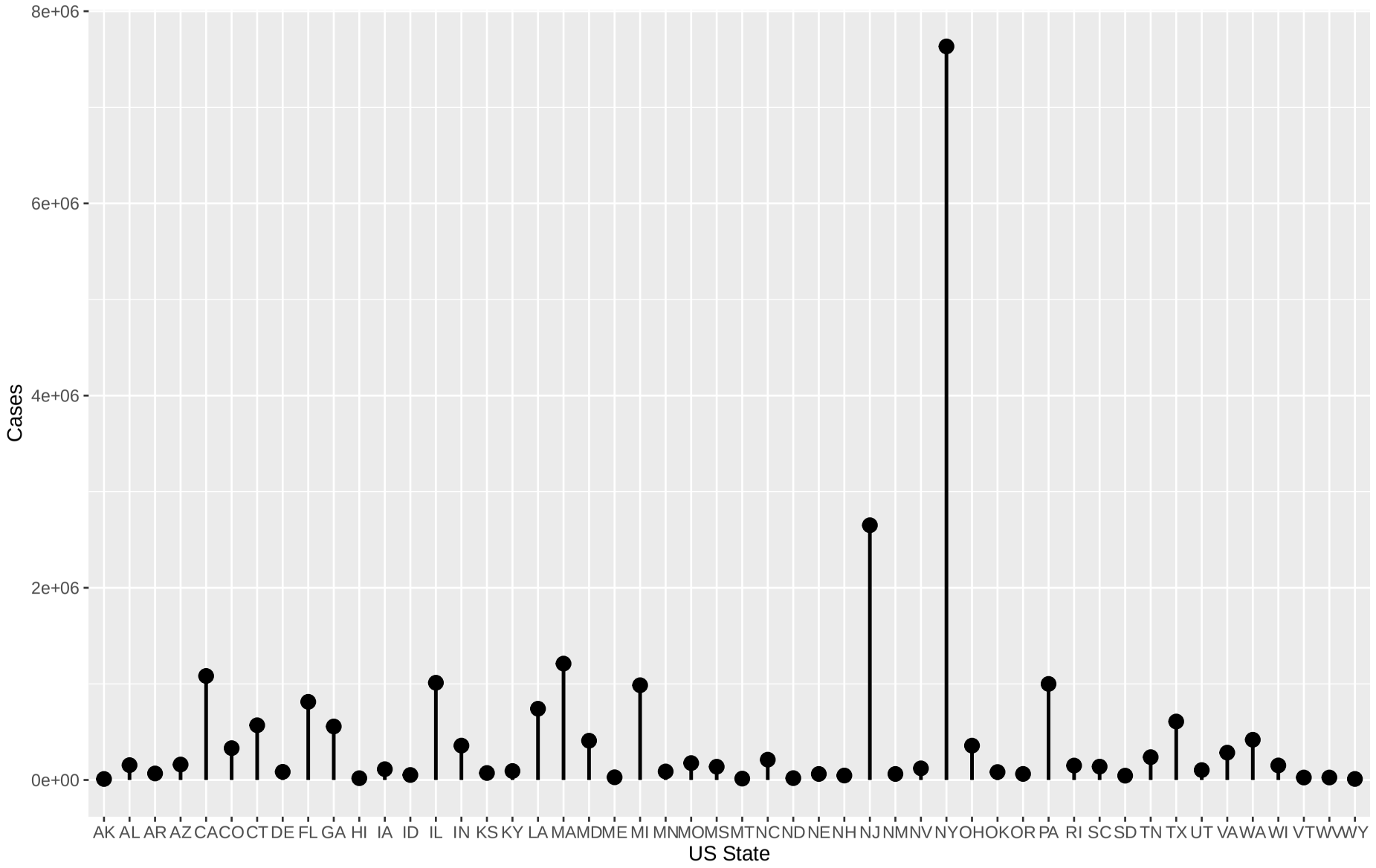

Figure 6: The number of Covid-19 deaths per state.