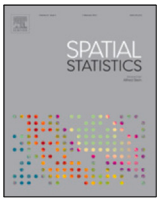
Spatial Statistics 49 (2022) 100504

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence

Francesco Bartolucci [a], Alessio Farcomeni [b,*]

[a] University of Perugia, Perugia, Italy
[b] University of Rome "Tor Vergata", via Columbia, 2, 00133 Roma, Italy

ARTICLE INFO

Article history:
Received 7 February 2021
Received in revised form 16 March 2021
Accepted 16 March 2021
Available online 27 March 2021

Keywords:
Data augmentation
Hidden Markov models
MCMC
SARS-CoV-2
Swabs

ABSTRACT

We propose a model based on discrete latent variables, which are spatially associated and time specific, for the analysis of incident cases of SARS-CoV-2 infections. We assume that for each area the sequence of latent variables across time follows a Markov chain with initial and transition probabilities that also depend on latent variables in neighboring areas. The model is estimated by a Markov chain Monte Carlo algorithm based on a data augmentation scheme, in which the latent states are drawn together with the model parameters for each area and time. As an illustration we analyze incident cases of SARS-CoV-2 collected in Italy at regional level for the period from February 24, 2020, to January 17, 2021, corresponding to 48 weeks, where we use number of swabs as an offset. Our model identifies a common trend and, for every week, assigns each region to one among five distinct risk groups.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

COVID-19 syndrome is due to newly discovered SARS-CoV-2 coronavirus, which acquired the ability to infect humans, and for human-to-human transmission, during 2019. See Hu et al. (2020) and references therein for a review.

Due to very limited pre-existing immunity, SARS-CoV-2 has the potential to be highly infectious. Simultaneously, pathogenicity is high enough to generate a proportion of severe cases (Buss et al., 2021; Del Sole et al., 2020) that can overwhelm health systems (Grasselli et al., 2020; Farcomeni et al., 2021a) when prevalence is high. Monitoring the epidemics is therefore a priority for policy, planning, and resource allocation.

* Corresponding author.
E-mail address: alessio.farcomeni@uniroma2.it (A. Farcomeni).

https://doi.org/10.1016/j.spasta.2021.100504
2211-6753/© 2021 Elsevier B.V. All rights reserved.

Public data for Italy are available at regional level. These include incident cases (i.e., new positives), prevalent cases (i.e., currently infected), deaths, number of infected currently in hospital wards, number of infected specifically hospitalized in Intensive Care Units, number of swabs, and number of tested cases. Data are timely updated, every day at around 6p.m., by the Italian Civil Protection Department. However, counts are subjected to measurement errors and biases of various nature, among which we mention two important sources of error that we take into account in our model. First of all, data collection is not standardized and, in many areas, not in electronic form. This results in frequent late notifications and communication errors that are corrected afterwards. Adjustments are made to the daily count, in order to obtain the correct cumulative number of cases; this may make daily counts inconsistent, and occasionally negative. This is the main reason why we will work with weekly incidence, even if counts are available on a daily basis. This is not ensured to resolve measurement issues but, as can be seen from our descriptive analyses, it mitigates them sufficiently. Secondly, it can be easily argued that the most prominent source of bias arises from undercount. Several infections from SARS-CoV-2 are asymptomatic or paucisymptomatic, and will easily go undetected (Li et al., 2020). This is linked to the difficulties in testing and tracing (Contreras et al., 2021). Indeed, SARS-CoV-2 infection at first could be detected only through polymerase chain reaction of nasopharingeal swabs, whose availability was very limited, and yet has not scaled sufficiently. Notably, in the first months of 2020, swabs were mostly used in Italy for confirmation of severe symptomatic cases, with the exception of Veneto and some provinces of other regions. Diagnostic testing was later extended also to asymptomatic subjects, through contact tracing and screening strategies. Contact tracing efforts have varied wildly over time and space due to unobserved reasons (e.g., availability of swabs, tracers, underlying incidence, policies), making the proportion of identified cases highly and unpredictably variable over time and space. This is another important reason why we will model weekly counts using swabs as an offset, essentially studying positive rate rather than incidence. We prefer to offset with respect to the number of swabs rather than that of tested cases, as the latter has been added to the data set only starting from April 19, 2020. More importantly, a positive rate defined with respect to the number of swabs has been identified by WHO as an official indicator of a nation's ability to flatten the curve.

Being focused on weekly positive rate, and on a flexible model that can take into account spatio-temporal unobserved heterogeneity, we claim that our analysis is somehow more reliable than more common analyses simply focused on fixed-effects models for daily incident cases. On the other hand, our results should be interpreted with some care: a high positive rate is an indication of inability to perform contact tracing, and only an indirect indication of high incidence.

In order to analyze data of the type described above, we propose a model based on discrete latent variables, so that regions may be clustered in groups corresponding to different levels of severity. More precisely, we adopt a model with a structural component formulated in the spirit of Bartolucci and Farcomeni (2021). Discrete latent variables are assumed to depend on the current status of their neighbors according to an auto-logistic model (Besag, 1974). More precisely, each area-time-varying latent distribution depends, via a logistic parameterization, on the latent states of the neighbors for the same time occasion, and on the previous latent state for the same site (as in a standard hidden Markov model). For each geographical region, the sequence of latent variables is assumed to follow a first-order Markov chain with initial and transition probabilities depending on the latent states of the neighboring areas. The model may be seen as an extension of a hidden Markov model for longitudinal data (Bartolucci et al., 2013, 2014), accounting for spatial interaction. From another perspective, we are adopting a temporal extension of a hidden Markov field model (Qian and Titterington, 1991; Green and Richardson, 2002; Spezia et al., 2018) for spatial data, based on discrete latent variables. Unlike Bartolucci and Farcomeni (2021), which is restricted to binary outcomes, in this work we outline a class of models based on general parametric assumptions on the outcome. The proposed inferential strategy seems to be stable under different model specifications. Another distinctive feature with respect to Bartolucci and Farcomeni (2021), that we find very useful to modeling COVID-19 data, is that a common trend is estimated through splines, therefore without strict parametric assumptions on its shape. This allows us to automatically fit more pandemic waves within the same model. On the other hand, more common parametric models, typically based on logistic growth curves (e.g., Cabras, 2020; Girardi et al., 2020; Alaimo Di Loro et al., 2020), are

restricted to modeling only a single wave at a time, and often involve arbitrarily setting an initial and final date for the specific wave. For model estimation we adopt a Markov chain Monte Carlo (MCMC) algorithm that extends the proposal of Bartolucci and Farcomeni (2021) to the model class here adopted. The algorithm is based on data augmentation (Tanner and Wong, 1987), and alternates different steps at which parameter values and latent states are drawn. For selecting the number of latent states we adopt a simple strategy based on post-processing the MCMC output, thus avoiding computational overheads and issues with respect to estimating the marginal likelihood or ratios of marginal likelihoods.

The particular model we use for COVID-19 data assumes that the number of incident cases for a certain area and time occasion follows a Poisson regression model, with offset equal to the logarithm of the number of swabs, conditionally on a linear predictor that depends on a specific discrete latent variable. In particular, the log-linear predictor is based on splines of time of suitable order common to all areas, and a shift that depends on the value of the underlying latent variable. This allows us to separate a common time trend from unobserved heterogeneity, which is completely captured by the latent variable. Additionally, our specification allows us to cluster areas with respect to shifts from the common trend, an operation that can be used to identify risk profiles in an unsupervised manner. Italian authorities have worked for some time with three main risk profiles (areas can be yellow, orange, or red) and have recently increased the number of risk profiles to five (adding white and "dark orange" groups). Risk profiles are currently identified according to an algorithm which is mainly based on estimates of effective reproductive number, restricted to the symptomatic cases.

The rest of the paper is organized as follows. In the following section we illustrate the assumptions of the proposed model. Bayesian inference, and in particular the MCMC algorithm for parameter estimation, is described in Section 3 together with model selection. The application to Italian regional data is described, together with the data structure, in Section 4. Section 5 provides some conclusions, and lists routes for possible extensions.

2. Model assumptions

Let Y_{it} be the count variable of interest for area i at time t, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. In our context this variable corresponds to the daily number of positives in a certain region. Let y_{it} denote the observed value of Y_{it}, with all values collected in the matrix $\boldsymbol{Y}$. Finally, we represent the spatial structure by the indicator variables c_{ij}, where $c_{ij} = 1$ if area j is in the neighborhood of i (and viceversa) and $c_{ij} = 0$ otherwise, for $i, j = 1, \ldots, n$.

The first assumption of our model is that a discrete latent variable U_{it} is associated to each area i and time t so that, given the set of all these latent variables, the response variables Y_{it} are conditionally independent. These latent variables have k support points, referred to as latent states, labeled from 1 to k. We assume that Y_{it}, conditionally on U_{it}, has a distribution belonging to the natural exponential family (see also McCullagh and Nelder, 1989; Farcomeni, 2015):

$$p(y_{it}|U_{it} = u, \boldsymbol{\eta}, \psi) = \exp\{[y_{it}\eta_{it}(u) - c(\eta_{it}(u))]/[a(\psi) - b(y_{it}, \psi)]\}, \tag{1}$$

where functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known, while $\eta_{it}(u)$ is a parameter of interest and ψ a nuisance parameter. We then adopt a specific link function $g(\cdot)$ and assume that

$$g(\eta_{it}(u)) = \xi_u + \boldsymbol{b}_t'\boldsymbol{\beta}, \tag{2}$$

where ξ_u is an intercept specific to latent state u and $\boldsymbol{b}_t$ is the base vector at time t of splines of a suitable order r, with prespecified knots (e.g., Wood, 2017). The parameters $\xi_1, \ldots, \xi_k$, together with $\boldsymbol{\beta}$, will collected in the vector $\boldsymbol{\phi}$.

In our application we assume that Y_{it}, conditional on U_{it}, follows a Poisson regression model with offset o_{it}. The offset is set equal to the logarithm of the number of swabs in the same area and for the same period. The log-linear predictor depends additionally on the value of the underlying latent variable, plus a common trend that is independent of the latent state. More formally, we assume that

$$Y_{it}|U_{it} = u \sim \text{Pois}(\lambda_{it}(u)), \quad i = 1, \ldots, n,\ t = 1, \ldots, T,\ u = 1, \ldots, k, \tag{3}$$

with

$$\log \lambda_{it}(u) = \xi_u + \boldsymbol{b}_t'\boldsymbol{\beta} + o_{it}. \tag{4}$$

Regarding the assumptions on the distribution latent variables, we slightly generalize Bartolucci and Farcomeni (2021) by allowing general dependence structures among each latent state and its neighbors, as we summarize in the following. Let $\tilde{\boldsymbol{u}}_{it}$ denote the vector of latent variables of the neighbors of area i at time t, that is, variables u_{jt} such that $c_{ij} = 1$. For the initial probabilities we assume that

$$\begin{aligned}\pi_i(u|\tilde{\boldsymbol{u}}_{i1}) &= p(U_{i1} = u|\tilde{\boldsymbol{u}}_{i1}) \\ &= \frac{1}{1 + \sum_{v=2}^{k} \exp(\boldsymbol{f}_1(\tilde{\boldsymbol{u}}_{i1})'\boldsymbol{\gamma}_v)} \begin{cases} 1, & u = 1, \\ \exp(\boldsymbol{f}_1(\tilde{\boldsymbol{u}}_{it})'\boldsymbol{\gamma}_u), & u = 2, \ldots, k, \end{cases}\end{aligned} \tag{5}$$

where $\boldsymbol{f}_1(\tilde{\boldsymbol{u}}_{it})$ is a known function. In our implementation we specify $\boldsymbol{f}_1(\cdot)$ so to obtain a vector with a leading unity (for the intercept) and the remaining elements corresponding to the proportion of neighbors currently dwelling in each latent state apart from the first. For ease of notation we collect $\boldsymbol{\gamma}_u$ vectors in the matrix $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_k)'$. Regarding the transition probabilities, we assume a similar parameterization, but using the starting state as reference category, that is,

$$\begin{aligned}\pi_{it}(u|u', \tilde{\boldsymbol{u}}_{it}) &= p(U_{it} = u|U_{i,t-1} = u', \tilde{\boldsymbol{u}}_{it}) \\ &= \frac{1}{1 + \sum_{\substack{v=1 \\ v \neq u}}^{k} \exp(\boldsymbol{f}_t(\tilde{\boldsymbol{u}}_{it})'\boldsymbol{\delta}_{u'v})} \begin{cases} 1, & u = u', \\ \exp(\boldsymbol{f}_t(\tilde{\boldsymbol{u}}_{it})'\boldsymbol{\delta}_{u'u}), & u \neq u', \end{cases}\end{aligned} \tag{6}$$

for $u', u = 1, \ldots, k$. Here $\boldsymbol{f}_t(\cdot)$ is a possibly time-dependent known function. In our implementation we specify $\boldsymbol{f}_t(\boldsymbol{u}) = \boldsymbol{f}_1(\boldsymbol{u})$ for all t. Parameters for the transition distribution are collected in matrix $\boldsymbol{\Delta}$, with vectors $\boldsymbol{\delta}_{u'u}$ for $u', u = 1, \ldots, k$, with $u \neq u'$.

For all parameters we assume prior independence and that they *a priori* follow a zero-centered Gaussian distribution. Formally, for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$ we assume that

$$\begin{aligned} p(\boldsymbol{\Gamma}) &= \prod_{u=2}^{k} p(\boldsymbol{\gamma}_u), \\ p(\boldsymbol{\Delta}) &= \prod_{u'=1}^{k} \prod_{\substack{u=1 \\ u \neq u'}}^{k} p(\boldsymbol{\delta}_{u'u}). \end{aligned}$$

3. Bayesian inference

In this section we describe an algorithm for approximately sampling from the posterior distribution, which is closely related to that in Bartolucci and Farcomeni (2021). We follow an augmentation scheme according to which latent states are sampled from their full conditional at each iteration. Let $\boldsymbol{U}$ denote a matrix with element u_{it} for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Formally, Bayes theorem can be used to show that the posterior distribution $p(\boldsymbol{\phi}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \boldsymbol{U}|\boldsymbol{Y})$ is proportional to

$$p(\boldsymbol{\phi})p(\boldsymbol{\Gamma})p(\boldsymbol{\Delta})p(\boldsymbol{U}|\boldsymbol{\Gamma}, \boldsymbol{\Delta})p(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{\phi}), \tag{7}$$

where the normalizing constant cannot be obtained in closed form.

A particularly challenging factor to be computed among those in (7) is the full conditional distribution of $\boldsymbol{U}$, considering the assumed spatio-temporal dependence structure. We approximate this distribution as in Bartolucci and Farcomeni (2021) through the following pseudo-probability

$$\tilde{p}(\boldsymbol{U}|\boldsymbol{\Gamma}, \boldsymbol{\Delta}) = \prod_{i=1}^{n} \left[\pi(u_{i1}|\tilde{\boldsymbol{u}}_{i1}, \boldsymbol{\Gamma}) \prod_{t=2}^{T} \pi(u_{it}|u_{i,t-1}, \tilde{\boldsymbol{u}}_{it}, \boldsymbol{\Delta}) \right].$$

It must be clarified that the above expression coincides with the true probability under spatial independence, whereas in general the quality of the approximation decreases as the spatial dependence

becomes stronger. However, adopting this approximation is rather common in spatial statistics even for simpler models, and leads to the definition of a pseudo-likelihood (Besag, 1975); see also Spezia et al. (2018). A brief additional discussion on this choice is given in the concluding section.

Regarding the conditional distribution of the response variables Y_{it}, with realizations collected in $\boldsymbol{Y}$, under the conditional independence assumption it is straightforward to see that

$$p(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\phi}) = \prod_{i=1}^{n}\prod_{t=1}^{T} p(y_{it}|u_{it},\boldsymbol{\phi}),$$

where $p(y_{it}|u_{it},\boldsymbol{\phi})$ is the density or probability mass function corresponding to the assumed distribution, that belongs to the natural exponential family, defined in (1). Dependence on $\boldsymbol{\phi}$ is formulated according to (2). For our specific application, each response variable has a conditional Poisson distribution defined as in (3) and link function defined in (4).

3.1. Markov chain Monte Carlo algorithm

In order to approximate the posterior distribution of model parameters, we rely on an MCMC algorithm, with an augmented parameter space. The algorithm is based on repeating a sequence of iterations for a large number of times R, where initial iterations are finally discarded (burn-in). Each of these steps can be summarized as follows:

- **Update of the $\boldsymbol{\phi}$ parameters:** A candidate update ϕ_j^* is sampled from a Gaussian distribution centered at the current value ϕ_j, with variance τ_ϕ^2. The proposed parameter vector is accepted with probability
$$\alpha(\boldsymbol{\phi}^*,\boldsymbol{\phi}) = \min\left(1, \frac{p(\boldsymbol{\phi}^*)p(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\phi}^*)}{p(\boldsymbol{\phi})p(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\phi})}\right).$$
- **Update of the latent variable values** u_{it}: here, differently from Bartolucci and Farcomeni (2021), for $i = 1,\dots,n$ and $t = 1,\dots,T$ we use a pseudo-Gibbs sampling step, based on the k-dimensional vector of probabilities $\boldsymbol{q}_{it}$ with elements equal to
$$q_{it}(u) = \frac{\tilde{p}(\boldsymbol{U}_{it}(u)|\boldsymbol{\Gamma},\boldsymbol{\Delta})p(\boldsymbol{Y}|\boldsymbol{U}_{it}(u),\boldsymbol{\phi})}{\sum_{v=1}^{k}\tilde{p}(\boldsymbol{U}_{it}(v)|\boldsymbol{\Gamma},\boldsymbol{\Delta})p(\boldsymbol{Y}|\boldsymbol{U}_{it}(v),\boldsymbol{\phi})}, \quad u = 1,\dots,k,$$
where $\boldsymbol{U}_{it}(u)$ is the current matrix of latent states $\boldsymbol{U}$, with elements u_{it} substituted by u. The new value of U_{it} is drawn from a Multinomial distribution with parameters 1 and $\boldsymbol{q}_{it}$.
- **Update of the $\boldsymbol{\gamma}_u$ parameters:** for $u = 2,\dots,k$ we propose a new parameter vector $\boldsymbol{\gamma}_u^*$ which is accepted with probability
$$\alpha(\boldsymbol{\gamma}_u,\boldsymbol{\gamma}_u^*) = \min\left(1, \frac{p(\boldsymbol{\gamma}_u^*)\prod_{i=1}^{n}\pi_i(u_{i1}|\tilde{\boldsymbol{u}}_{i1};\boldsymbol{\Gamma}_u^*)}{p(\boldsymbol{\gamma}_u)\prod_{i=1}^{n}\pi_i(u_{i1}|\tilde{\boldsymbol{u}}_{i1};\boldsymbol{\Gamma})}\right),$$
where $\boldsymbol{\Gamma}_u^*$ is the current matrix $\boldsymbol{\Gamma}$ with vector $\boldsymbol{\gamma}_u$ substituted by $\boldsymbol{\gamma}_u^*$.
- **Update of the $\boldsymbol{\delta}_{u'u}$ parameters:** for $u', u = 1,\dots,k$, with $u \neq u'$, we propose a new parameter vector $\boldsymbol{\delta}_{u'u}$, denoted by $\boldsymbol{\delta}_{u'u}^*$, which is accepted with probability
$$\alpha(\boldsymbol{\delta}_{u'u},\boldsymbol{\delta}_{u'u}^*) = \min\left(1, \frac{p(\boldsymbol{\delta}_{u'u}^*)\prod_{i=1}^{n}\prod_{t=2}^{T}\pi_{it}(u_{it}|u_{i,t-1},\tilde{\boldsymbol{u}}_{it};\boldsymbol{\Delta}_{u'u}^*)}{p(\boldsymbol{\delta}_{u'u})\prod_{i=1}^{n}\prod_{t=2}^{T}\pi_{it}(u_{it}|u_{i,t-1},\tilde{\boldsymbol{u}}_{it};\boldsymbol{\Delta})}\right),$$
where $\boldsymbol{\Delta}_{u'u}^*$ is the matrix $\boldsymbol{\Delta}$ with parameter vector $\boldsymbol{\delta}_{u'u}$ substituted by $\boldsymbol{\delta}_{u'u}^*$. The products in the previous expression may be restricted to the only cases in which $u_{i,t-1} = u'$.

The MCMC output may be elaborated in the usual way to obtain point estimates and credible intervals for the model parameters. Similarly, latent states are predicted according to a *maximum a posteriori* rule, that is, the predicted latent state for each area at each time point corresponds to the latent state most frequently sampled during the MCMC iterations, after ignoring burn-in.

It is straightforward to see that the model is label independent and that consequently the posterior distribution will have $k!$ modes. There are several ways of dealing with this label switching issue. In our implementation we use an on-line approach that, at every iteration of the MCMC algorithm, maps the sampled parameters so that latent intercepts $\xi_1, \ldots, \xi_k$ are increasingly ordered. This is simpler and implies a reduced computational load with respect to the post-processing approach of Bartolucci and Farcomeni (2021).

For selecting the number of support points of the latent variables we use a similar approach to Bartolucci and Farcomeni (2021). In our Bayesian framework it would be natural to set up a reversible jump sampling scheme, after specification of a prior distribution for the number of latent states. This would anyway be cumbersome both from a computational and formal perspective, as acceptance probabilities for transdimensional moves are not easily derived under the current general model formulation. A simple solution is to repeatedly fit the model for different values of k, and compare the results through an appropriate tool like an information criterion. This is particularly advantageous from a computational perspective, as parallel computing can be set up in order to simultaneously fit the model for different number of latent states. In particular, we suggest to rely on the WAIC (Watanabe, 2010; Vehtari et al., 2017), which is a measure of the predictive accuracy and is a direct by-product of our MCMC sampling scheme. For each model specification, the WAIC is computed first by evaluating the log-pointwise predictive density

$$\widehat{\text{lpd}} = \sum_{i=1}^{n}\sum_{t=1}^{T} \log\left(\frac{1}{R}\sum_{r=1}^{R} p(y_{it}|\boldsymbol{\theta}^{(r)})\right)$$

and

$$\hat{p}_{\text{waic}} = \sum_{l=1}^{n}\sum_{t=1}^{T} \hat{V}(\log p(y_{it}|\boldsymbol{\theta})),$$

where $\hat{V}(\log p(y_{it}|\boldsymbol{\theta}))$ is the variance of $\log p(y_{it}|\boldsymbol{\theta}^{(r)})$ across the R iterations producing parameter vectors $\boldsymbol{\theta}^{(r)}$. The expected log-pointwise predictive density for a new data set is then computed as

$$\widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{lpd}} - \hat{p}_{\text{waic}}. \tag{8}$$

Due to a certain instability in the estimation of this variance that we noted in the application, the above quantities are computed only using the final part of the MCMC iterations. An advantage of the WAIC is that it does not involve the marginal likelihood, and it is simpler to compute than other Bayesian criteria. Additionally, unlike information criteria adopted in the frequentist approach, it does not require computation of the maximum of the observed likelihood.

4. Application

We applied the approach illustrated in the previous sections to the Italian regional data for the period from February 24, 2020, to January 17, 2021. There are wide daily oscillations and a clear weekly seasonality, with substantially fewer swabs on Sunday with respect to the rest of the week. In order to overcome problems with data reporting we simply aggregate data at weekly level, as already mentioned. In the end, we record $T = 48$ weeks of observation for each of $n = 21$ areas (19 corresponding to Italian regions and 2 corresponding to the two provinces of the Trentino Alto Adige region: Bolzano and Trento).

The analysis is based on the number of new positives reported in each week, and the corresponding number of swabs. The first is our outcome of interest y_{it}, while, on the basis of the second, we obtain the offset o_{it}. Our aims with this analysis are to: (*i*) identify a common trend for Italy, after having taken into account variation at regional level that is due to time-varying unobserved heterogeneity and spatial dependence; (*ii*) compare regional trends with respect to the common trend; (*iii*) identify latent trajectories in order to identify different risk profiles; and (*iv*) identify latent clusters in order to dynamically assign (i.e., specifically for each week) a region to a risk profile.

Table 1
Descriptive statistics for the number of swabs, number of positives, and rate of positives at regional and Italian levels. P. A. stands for "Provincia Autonoma".

Region	Area	N. swabs			N. positives			Rate		
		min	mean	max	min	mean	max	min	mean	max
Abruzzo	South	52	12 104	31 584	5	838	4 666	0.0009	0.0538	0.2301
Basilicata	South	39	4 274	12 729	0	267	1 591	0.0000	0.0426	0.1584
Calabria	South	35	10 244	23 170	1	619	3 803	0.0001	0.0423	0.1641
Campania	South	380	47 614	164 936	8	4 411	27 319	0.0003	0.0615	0.2209
Emilia-Romagna	North	1 795	59 863	131 603	118	4 302	17 218	0.0032	0.0741	0.3449
Friuli Venezia Giulia	North	243	22 081	55 090	3	1 297	5 721	0.0002	0.0423	0.1297
Lazio	Center	724	63 045	188 071	5	4 035	18 460	0.0030	0.0436	0.1350
Liguria	North	135	16 482	41 176	32	1 400	7 086	0.0036	0.0891	0.3660
Lombardia	North	7 422	111 339	293 848	391	10 940	60 026	0.0072	0.0958	0.4609
Marche	Center	103	13 161	41 775	8	1 066	4 527	0.0012	0.0837	0.4482
Molise	South	6	2 707	7 928	0	161	914	0.0000	0.0450	0.1701
P. A. Bolzano	North	21	8 616	26 080	1	714	4 210	0.0007	0.0648	0.3810
P. A. Trento	North	116	10 363	22 639	0	525	1 740	0.0000	0.0584	0.3737
Piemonte	North	362	43 306	145 671	38	4 582	27 686	0.0021	0.0932	0.3974
Puglia	South	262	25 143	63 646	3	2 328	11 123	0.0003	0.0612	0.1875
Sardegna	South	29	11 225	28 203	0	762	3 310	0.0000	0.0511	0.1465
Sicilia	South	295	30 588	134 349	5	2 577	12 674	0.0003	0.0549	0.1764
Toscana	Center	572	43 394	125 867	13	2 720	16 457	0.0015	0.0477	0.1840
Umbria	Center	35	11 718	30 907	1	687	3 992	0.0001	0.0448	0.1805
Valle d'Aosta	North	10	1 567	5 019	0	163	1 111	0.0000	0.0857	0.4264
Veneto	North	6 862	77 527	177 645	26	6 344	26 282	0.0004	0.0660	0.2263
North		19 166	351 144	787 733	1034	30 267	144 572	0.0050	0.0775	0.2931
Center		1 434	131 318	360 265	45	8 509	43 381	0.0019	0.0497	0.1737
South		1 098	143 899	388 914	35	11 963	59 837	0.0006	0.0557	0.1568
Italy		21 698	626 361	1 510 190	1306	50 738	243 444	0.0040	0.0687	0.2543

4.1. Data description

In Table 1 we report descriptive statistics about the number of swabs, number of positives, and rate of positives at regional and Italian levels. We observe heterogeneity among regions, with an average rate of positives that goes from 4.23% for Calabria to 9.58% for Lombardia, while the Italian average rate is 6.87%. Regarding the temporal pattern, in Figs. 1 and 2 we represent the weekly positive rate region by region, together with the Italian tendency obtained by fitting our model with $k = 1$, that is, a fixed-effect Poisson spline regression with offset. This regression model is based on splines of cubic order, and knots at each fourth week starting from the fifth.

In terms of partition between areas, the North presents an average rate of 7.75%, which is distant from that of the Center and South that have a rate around 5%. This is due to the fact that northern regions, and especially Lombardia, were hit very hard during the first outbreak. The limited amount of swabs available made it very difficult to perform contact tracing or screening, and for some time swabs were reserved to symptomatic cases with a high index of suspicion, resulting in high positive rates. This is apparent not only for Lombardia but also for Liguria, that experienced a comparatively large outbreak in the late Summer of 2020. A high heterogeneity is also observed in the Center, where the Marche region has an average positive rate of 8.37%, much higher than the other central regions, whereas less heterogeneity is observed in the South. All regions present two or three peaks in the temporal distribution of positive rates, with Lombardia, Marche, and Piemonte clearly being hit harder than other regions in the first few weeks. An interesting example is that of Veneto, which managed well the first wave through relatively massive testing, but then had troubles in Autumn 2020. It is speculated that this is due to the somehow restricted use of molecular swabs to confirm results of rapid antigen tests, which are less accurate.

The adjacency matrix has been defined by considering regions sharing land borders as neighbors. As a consequence, seven regions have three neighbors, which is the mode. Two regions have six

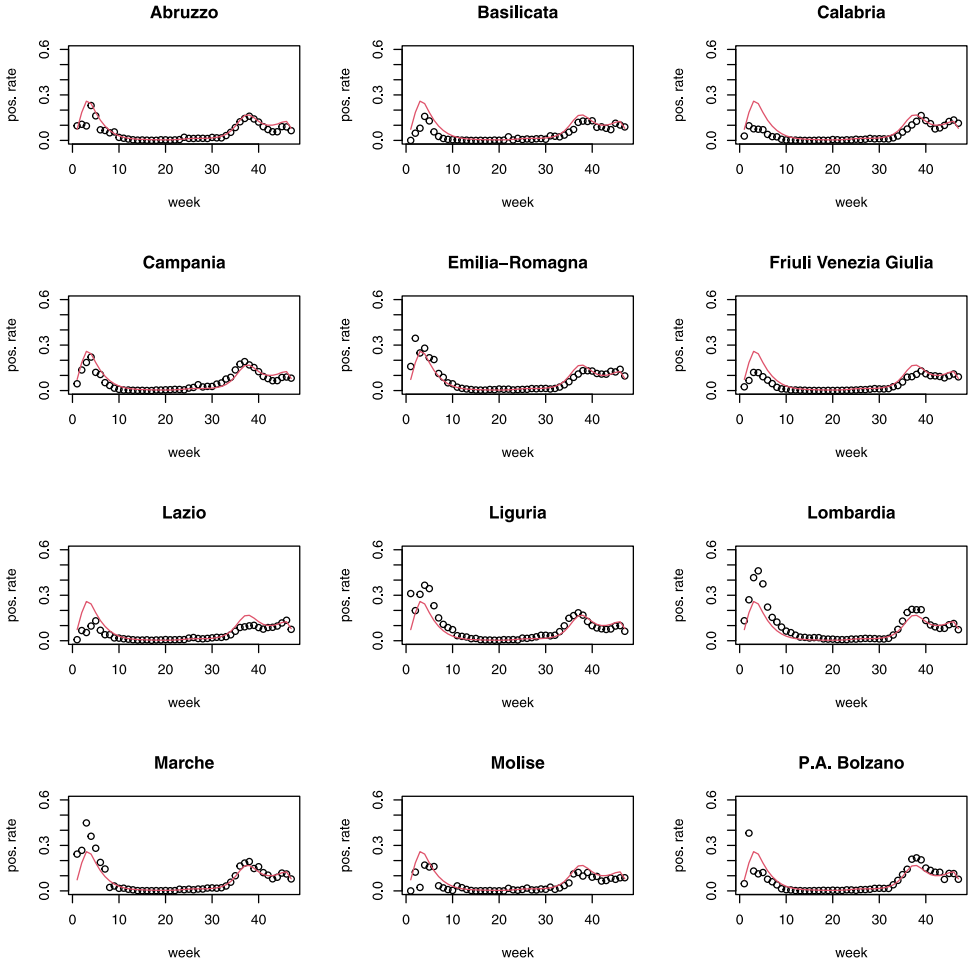


Fig. 1. Positive rate for twelve regions, together with the Italian tendency (in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

neighbors. Sardinia, which is an island, does not have any neighbor according to our definition. For the other island, Sicily, we selected Calabria as the only neighbor. The two regions are separated by the few kilometers wide Strict of Messina, with ferries going back and forth more than 150 times a day.

In Table 2 we report the average positive rate for each region in comparison with that of the neighborhood, in order to better appraise the spatial pattern. The correlation between these two quantities is equal to 0.589, which is in agreement with the heterogeneity of the regional situations already noted in commenting Table 1.

4.2. Data analysis

In applying our approach we considered the model based on the splines and defined by Eqs. (3) and (4). In particular, as already mentioned, we considered splines of cubic order with knots every four weeks starting from the fifth, the same adopted to describe the national Italian pattern in the previous section.

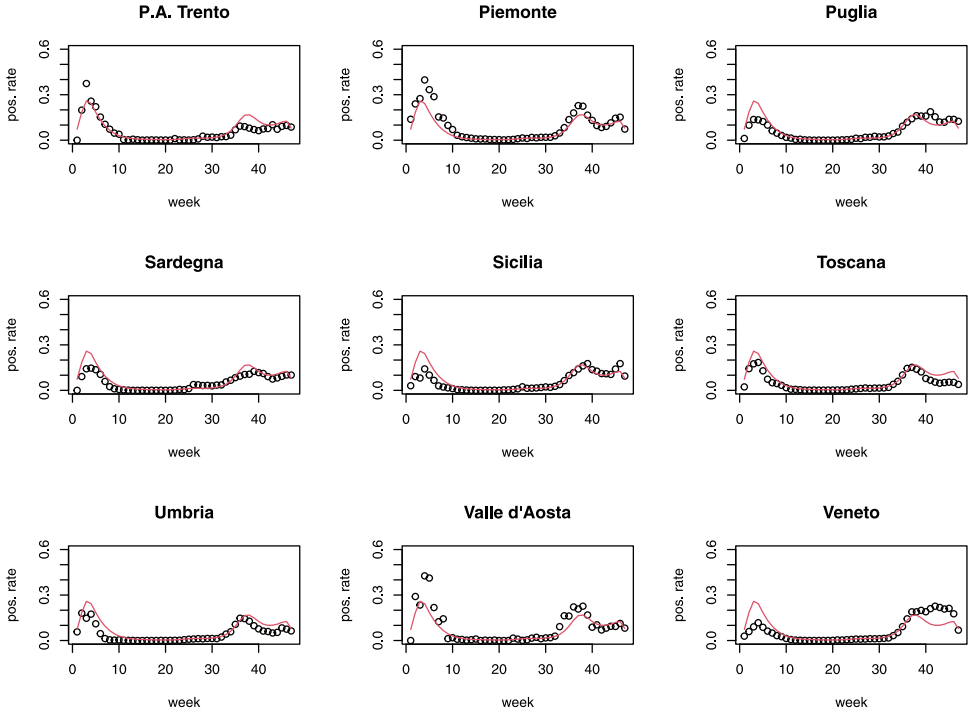


Fig. 2. Positive rate for nine regions, together with the Italian tendency (in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Number of neighbors, average positive rate, and positive rate of the neighborhood for each region.

Region	N. neighbors	Mean rate	
		Region	Neighborhood
Abruzzo	3	0.0538	0.0519
Basilicata	3	0.0426	0.0579
Calabria	2	0.0423	0.0533
Campania	4	0.0615	0.0477
Emilia-Romagna	6	0.0741	0.0780
Friuli Venezia Giulia	1	0.0423	0.0660
Lazio	6	0.0436	0.0579
Liguria	3	0.0891	0.0726
Lombardia	5	0.0958	0.0690
Marche	5	0.0837	0.0569
Molise	4	0.0450	0.0518
P. A. Bolzano	3	0.0648	0.0803
P. A. Trento	3	0.0584	0.0805
Piemonte	4	0.0932	0.0884
Puglia	3	0.0612	0.0590
Sardegna	0	0.0511	–
Sicilia	1	0.0549	0.0423
Toscana	5	0.0477	0.0645
Umbria	3	0.0448	0.0503
Valle d'Aosta	1	0.0857	0.0932
Veneto	5	0.0660	0.0819

Table 3
Results in terms of $\widehat{\text{elpd}}_{\text{waic}}$ and corresponding standard deviation for the model based on splines for different values of the number of latent states k.

u	$\widehat{\text{elpd}}_{\text{waic}}$	$\text{se}(\widehat{\text{elpd}}_{\text{waic}})$
1	−169 866.03	15 747.37
2	−66 676.26	4 219.82
3	−35 660.02	2 096.13
4	−23 159.70	1 355.10
5	−19 690.56	1 201.52
6	−17 079.82	1 041.95

Table 4
Estimated intercepts of each latent state.

u	$\hat{\xi}_u$	$\text{se}(\hat{\xi}_u)$
1	−3.425	0.0100
2	−3.057	0.0114
3	−2.842	0.0130
4	−2.613	0.0126
5	−2.130	0.0112

For this model we considered a number of latent states (k) from 1 to 6. Each model was estimated by an MCMC algorithm based on 10^5 iterations after a burnin of 2.5×10^4 iterations. In order to reduce the autocorrelation between consecutive draws, we fixed a thinning of 100. The convergence diagnostics are good for all values of k, with low autocorrelation functions for the final chains. The acceptance rates for the Metropolis-within-Gibbs steps are sensible for all parameters.

The results from the MCMC algorithm in terms of $\widehat{\text{elpd}}_{\text{waic}}$, as defined in (8), and its standard deviation, are reported in Table 3. On the basis of theese results, taking into account also the standard deviation, we selected the model based on $k = 5$ latent states. This result is particularly interesting since authorities have recently switched from a three-level to a five-level regime (white, yellow, orange, dark orange, red). Our results suggest a mild evidence that, over time, there might actually have been already five different risk levels to differentiate regions with respect to a common trend.

In order to interpret our risk stratification, we report in Table 4 posterior summaries for the latent intercepts. We note that latent intercepts are roughly equally spaced, and that posterior distributions seem to be well separated. Jointly with the estimate of the regression coefficients in $\boldsymbol{\beta}$, which are not reported here because their interpretation is not straightforward, we obtain five trajectories that are represented in Fig. 3.

The five latent states correspond to increasing degrees of severity in terms of number of positives, conditionally on the number of swabs. The corresponding trajectories are represented in Fig. 3 in comparison with the Italian trend directly obtained from the observed data, the same used in Figs. 1 and 2. We observe that there are two trajectories that are uniformly below the national trend (for states 1–2) and two that are uniformly above the national trend (for states 4–5). The third state corresponds to a trend very close to the national one, but not uniformly above or below the latter.

It is important to consider that each region may be in one of the five latent states at each time occasion, and many regions move among latent states during the observation period. Each region moves between these trajectories according to our hidden Markov formulation, depending on the parameters $\boldsymbol{\gamma}_u$ and $\boldsymbol{\delta}_{u'u}$ in (5) and (6), respectively. Rather than reporting the estimates of these parameters, to favor interpretability we directly illustrate the estimated latent structure by reporting the distribution of the predicted latent states. In fact, the MCMC algorithm also allows us to assign each region to a latent state in a dynamic fashion, on the basis of the number of visits to these states. The posterior distribution of the latent states for each region, across weeks, is reported in Table 5. The time variation of assigned latent states, in each week, is shown in a heatmap in Fig. 4. A clear pattern emerges for most regions, which we better comment below. It shall be noted that

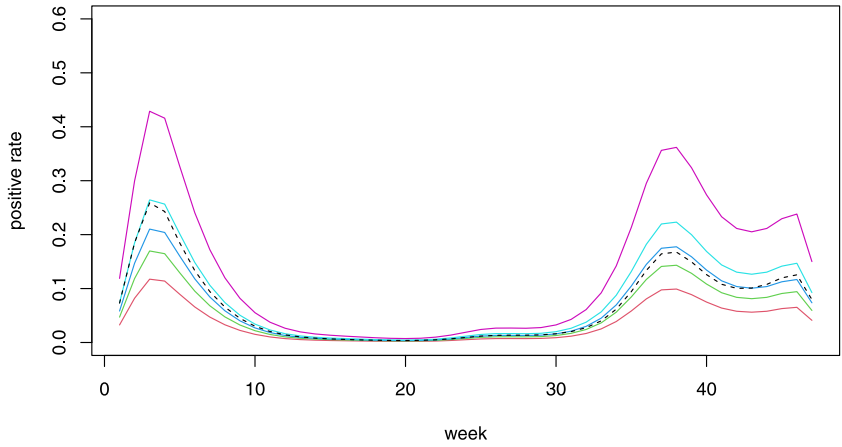


Fig. 3. Trajectories corresponding to the 5 latent states suitably ordered together with the Italian trend (dashed curve).

Table 5
Distribution of the predicted latent states across weeks at regional level.

Region	Latent state				
	1	2	3	4	5
Abruzzo	0.383	0.404	0.170	0.021	0.021
Basilicata	0.617	0.191	0.085	0.064	0.043
Calabria	0.702	0.085	0.106	0.085	0.021
Campania	0.362	0.149	0.170	0.106	0.213
Emilia-Romagna	0.234	0.128	0.191	0.255	0.191
Friuli Venezia Giulia	0.745	0.149	0.085	0.021	0.000
Lazio	0.489	0.106	0.170	0.170	0.064
Liguria	0.043	0.213	0.064	0.319	0.362
Lombardia	0.043	0.106	0.234	0.170	0.447
Marche	0.298	0.213	0.277	0.043	0.170
Molise	0.553	0.298	0.043	0.043	0.064
P. A. Bolzano	0.426	0.128	0.234	0.191	0.021
P. A. Trento	0.511	0.170	0.021	0.213	0.085
Piemonte	0.021	0.106	0.234	0.383	0.255
Puglia	0.383	0.213	0.149	0.191	0.064
Sardegna	0.468	0.298	0.043	0.064	0.128
Sicilia	0.426	0.128	0.277	0.149	0.021
Toscana	0.553	0.277	0.149	0.021	0.000
Umbria	0.596	0.277	0.085	0.021	0.021
Valle d'Aosta	0.191	0.319	0.191	0.106	0.191
Veneto	0.532	0.213	0.085	0.064	0.106
Italy	0.408	0.199	0.146	0.129	0.119

Valle D'Aosta is one of the few regions without a clear emerging temporal pattern. We speculate this is due to the high variability of positive rate over time related to the reduced number of tests and positives in this small region. As a result, identification of small clusters could have an observable effect on the positive rate in this region. We additionally report in Table 6 the predicted latent states for each region across three periods, which define as the first wave (first fifteen weeks, until beginning of June, 2020), the transition period (the following eleven weeks, until mid-August), and the second wave (the remaining observation period).

Pooling information across regions, it can be argued that the latent state corresponding to the lowest risk is the most visited (40.8% of time-area occasions), whereas that corresponding to the highest risk is visited 11.9% of time-area occasions. The distribution of the latent states at regional

Table 6
Distribution of the predicted latent states across epidemic phases, at regional level.

Region	First wave Latent state					Transition time Latent state					Second wave Latent state				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Abruzzo	0.47	0.20	0.13	0.13	0.07	0.36	0.09	0.36	0.09	0.09	0.00	0.18	0.50	0.32	0.00
Basilicata	0.80	0.20	0.00	0.00	0.00	0.82	0.00	0.00	0.09	0.09	0.14	0.14	0.50	0.18	0.05
Calabria	0.93	0.07	0.00	0.00	0.00	0.82	0.18	0.00	0.00	0.00	0.14	0.09	0.41	0.32	0.05
Campania	0.47	0.33	0.13	0.07	0.00	0.27	0.27	0.09	0.27	0.09	0.00	0.09	0.18	0.32	0.41
Emilia-Romagna	0.07	0.07	0.27	0.40	0.20	0.18	0.36	0.00	0.36	0.09	0.00	0.18	0.32	0.50	0.00
Friuli Venezia Giulia	0.53	0.47	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.05	0.23	0.50	0.23	0.00
Lazio	0.20	0.73	0.07	0.00	0.00	0.00	0.45	0.00	0.55	0.00	0.00	0.09	0.82	0.09	0.00
Liguria	0.00	0.00	0.13	0.13	0.73	0.09	0.09	0.55	0.18	0.09	0.00	0.00	0.45	0.18	0.36
Lombardia	0.00	0.00	0.00	0.00	1.00	0.00	0.09	0.36	0.09	0.45	0.00	0.14	0.41	0.23	0.23
Marche	0.27	0.27	0.00	0.00	0.47	0.64	0.18	0.09	0.09	0.00	0.00	0.09	0.32	0.45	0.14
Molise	0.53	0.00	0.00	0.40	0.07	0.73	0.00	0.09	0.00	0.18	0.09	0.18	0.36	0.36	0.00
P. A. Bolzano	0.47	0.27	0.20	0.00	0.07	0.64	0.18	0.00	0.18	0.00	0.00	0.14	0.27	0.45	0.14
P. A. Trento	0.27	0.13	0.07	0.53	0.00	0.91	0.00	0.00	0.00	0.09	0.05	0.27	0.50	0.09	0.09
Piemonte	0.00	0.00	0.13	0.27	0.60	0.00	0.18	0.55	0.27	0.00	0.00	0.00	0.32	0.45	0.23
Puglia	0.20	0.60	0.20	0.00	0.00	0.82	0.00	0.18	0.00	0.00	0.00	0.00	0.14	0.64	0.23
Sardegna	0.53	0.33	0.13	0.00	0.00	0.64	0.18	0.09	0.00	0.09	0.00	0.00	0.36	0.45	0.18
Sicilia	0.67	0.33	0.00	0.00	0.00	0.45	0.36	0.09	0.00	0.09	0.00	0.00	0.18	0.73	0.09
Toscana	0.33	0.47	0.20	0.00	0.00	0.73	0.18	0.09	0.00	0.00	0.00	0.41	0.32	0.27	0.00
Umbria	0.67	0.13	0.13	0.00	0.07	0.91	0.00	0.09	0.00	0.00	0.00	0.32	0.45	0.23	0.00
Valle d'Aosta	0.33	0.07	0.27	0.00	0.33	0.45	0.09	0.27	0.00	0.18	0.00	0.05	0.32	0.32	0.32
Veneto	0.53	0.47	0.00	0.00	0.00	0.45	0.55	0.00	0.00	0.00	0.00	0.27	0.18	0.09	0.45

level is in agreement with the data description in Section 4.1. Many regions are never or very rarely assigned to the last latent state whereas other regions, such as Liguria and Lombardia, are frequently assigned to this state and at the same time are rarely assigned to the latent state corresponding to the lowest risk. Regions that are assigned to the last state at least 10% of times (about equal to the national average) are in the North of Italy, with the exception of Marche and Campania. On the contrary, regions that are assigned at least 40% of times (corresponding about to the national average) to the first state are in the Center and South, with the exception of Friuli Venezia Giulia, Veneto, and Trentino provinces.

The temporal pattern in Fig. 4 is also of interest, and can be compared with the observed patterns in Figs. 1 and 2. There are regions that started in an unfavorable situation, namely in the last latent state, and then improved much towards the end of the period of observation, such as Lombardia and Liguria. Other regions have an opposite trend, with a clear worsening of the situation across time, such as Veneto and Friuli Venezia Giulia. More mixed trends are typically observed for regions in the Center and the South of Italy. For instance, Toscana and Umbria moved towards states of greater severity in an intermediate period, but then saw a decreased incidence. This is especially true for Toscana, after the end of a localized lockdown that lasted few weeks.

Finally, in order to illustrate the spatial patterns in terms of predicted latent states, in Table 7 we report a measure of agreement between the state predicted for each region and the states predicted for its neighbors. This index is obtained by counting, for each week, the number of neighbors having a predicted state equal to that of the region. Once collapsed over the time occasions, this count is divided by the number of weeks and that of neighbors. We also include the same agreement index computed across macroregions. The average of the agreement measure at Italian level is 0.361, which compares with a theoretical value of 0.2 in case of absence of spatial dependence, being $k = 5$ the number of latent states. In this regard there is a certain heterogeneity, with some regions showing a strong spatial dependence, such as Calabria, and others showing a weaker spatial dependence, such as Emilia-Romagna. Even when we aggregate the index at a macroregional level it can be noted that northern regions show a larger spatial dependence than Center and South. We speculate that strong spatial dependence could be due to ongoing pendolarism across regions even during lockdown periods, where people were indeed allowed to move across regions for work. Note,

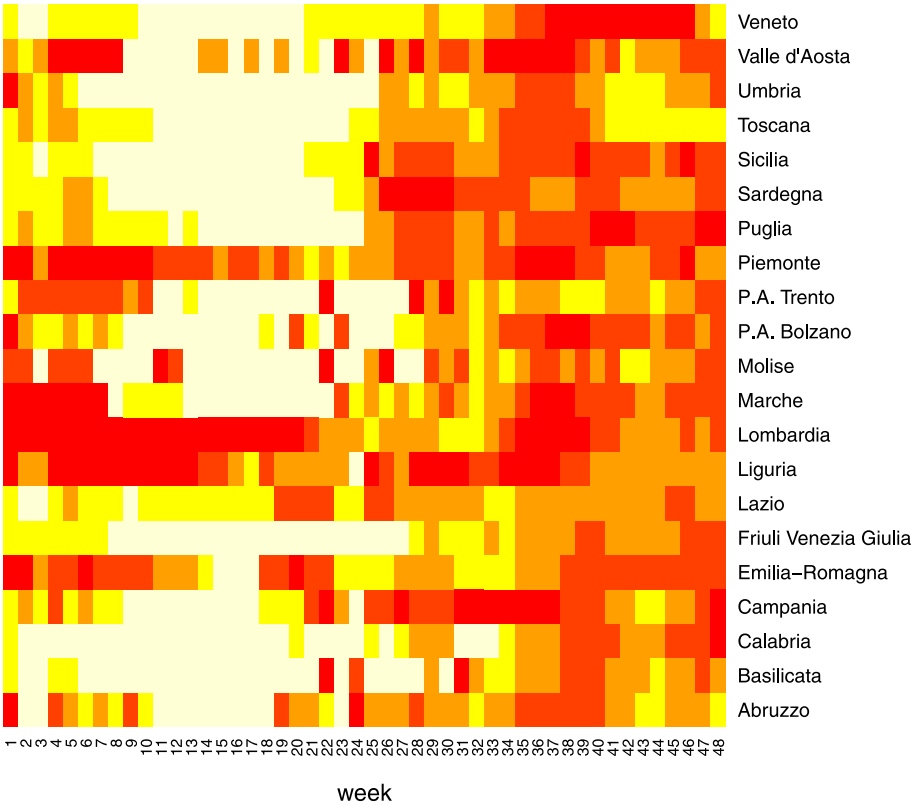


Fig. 4. Heatmap of the predicted latent state for each region at each week. Darker is associated with worse predicted risk.

however, that strength of spatial dependence might be influenced by the size of the neighborhood even if we divide this simple index by the number of neighbors. Overall, spatial dependence is far from irrelevant and it should then be appropriately modeled when analyzing the data at hand. This conclusion is also in agreement with what already noted in Section 4.1 on the basis of the results in Table 2.

5. Discussion

We propose a spatio-temporal approach for the analysis of weekly COVID-19 data, in which the number of swabs is used as an offset. Conditionally on a discrete latent variable, incident cases are assumed to follow a Poisson distribution modulated by a common trend, which is flexibly estimated using a spline model on the log-scale, and a multiplicative shock that depends on the latent state. Latent states evolve over time according to an inhomogeneous first-order Markov chain, so that areas can move from one level of risk to another. Levels of risk are therefore not absolute, but defined as proportional variations with respect to the common trend. Finally, spatial dependence is taken into account through the latent state, which depends on the state of neighboring areas at the same time occasion.

In contrast to other works using splines to approximate the emission densities (e.g., Langrock et al. (2015)), in this work we have specified them to describe a non-linear relationship between expected counts and time, in the spirit of Langrock et al. (2017). For ease of computation we have approximated the conditional distribution of the latent states through a pseudo-probability. This is

Table 7
Agreement between the state assigned to a certain region and to its neighbors.

Region	N. neighbors	Agreement
Abruzzo	3	0.326
Basilicata	3	0.482
Calabria	2	0.617
Campania	4	0.362
Emilia-Romagna	6	0.206
Friuli Venezia Giulia	1	0.574
Lazio	6	0.305
Liguria	3	0.213
Lombardia	5	0.255
Marche	5	0.306
Molise	4	0.372
P. A. Bolzano	3	0.355
P. A. Trento	3	0.234
Piemonte	4	0.351
Puglia	3	0.340
Sardegna	0	–
Sicilia	1	0.553
Toscana	5	0.311
Umbria	3	0.433
Valle d'Aosta	1	0.298
Veneto	5	0.332
North	3	0.134
Center	5	0.085
South	2	0.108
Italy	3	0.361

not uncommon in spatial statistics, and an excellent fit is obtained, even if some care might have to be used in interpreting parameters linked to the latent distribution. Readers are pointed to Friel and Pettitt (2004), Friel et al. (2009) and Everitt (2012) for further discussion on this point, and to Spezia et al. (2017) and references therein for other examples in spatial statistics. An alternative to this approximation would have been to use a nested sampling algorithm, according to which $p(\boldsymbol{U}|\Gamma, \boldsymbol{\Delta})$ is approximated at each MCMC iteration. This more rigorous approach would have been problematic from a computational point of view, as approximating the correctly specified full conditional for $\boldsymbol{U}$ would have implied a full-length MCMC algorithm within each outer MCMC iteration.

We have used the logarithm of the number of swabs as an offset in order to partially overcome bias due to the unknown and spatio-temporal heterogeneous sampling ratio. Our analysis, combined with weekly aggregation, gives a somehow robust assessment of five risk profiles and a common trend.

In our implementation we have defined a spatial structure that depends on sharing a land border. This is reasonable, but results should be interpreted considering this choice. Other choices are possible, such as defining a spatial structure on the basis of direct train or flight connections as in Della Rossa et al. (2020). While adopting a different spatial structure would probably lead to slightly different results regarding the distribution of latent states, we are confident that as long as enough and appropriate connections are specified, the goodness-of-fit and qualitative conclusions would be similar. The advantage of using our spatial structure is that it promotes similar latent states in adjacent regions, allowing authorities to act homogeneously in neighboring areas (e.g., specifying a policy for all regions in the north-east of Italy rather than for regions that are well connected from an economic and social point of view, but far apart from each other).

We leave to further work the derivation of a formal method for obtaining a posterior distribution on the number of latent states. One could also extend the model in order to allow for a time-varying number of latent states, as in Anderson et al. (2019), thus obtaining a different number of risk profiles at different times. Other possible extensions of our model involve use of more flexible parametric assumptions for the counts, for instance a negative Binomial, which would take

into account residual overdispersion, and extension to multivariate outcomes (e.g., a joint model for incident cases, hospital admissions, and deaths). While this would be straightforward using conditional independence assumptions (e.g., Bartolucci and Farcomeni, 2009, 2015; Farcomeni et al., 2021b), in our case the outcomes would have constraints that it is not straightforward to take into account. For instance, the (cumulative) number of deaths clearly cannot exceed the cumulative number of incident cases, and similarly for hospital admissions. Current assumptions, also, allow us to identify additive unobserved effects with respect to a common trend, essentially resulting in a multiplicative shift of the same. Much more care would be needed to identify a different trend for each latent state. Another possible extension would be the use of covariates at site level, which would allow us for instance to catch cyclic weekly effects (if modeling daily counts). Covariates at site level might also include indicators of interventions (like lockdowns, curfews, school closures), but interpretation of effects would require a lot of care due to endogeneity.

Finally, we believe the methodological device we put forward in this work is not only useful for analysis of COVID-19 data, but it can be applied with minor changes also in other areas of disease mapping. In applications of the latent Markov framework, indeed, spatial dependence might often be present and is generally ignored (e.g., Dotto et al., 2019).

Acknowledgments

The authors are grateful to three anonymous reviewers for constructive comments.

References

Alaimo Di Loro, P., Divino, F., Farcomeni, A., Jona Lasinio, G., Lovison, G., Maruotti, A., Mingione, M., 2020. Nowcasting COVID-19 incidence indicators during the Italian first outbreak. arXiv:2010.12679.

Anderson, G., Farcomeni, A., Pittau, M.G., Zelli, R., 2019. Rectangular latent Markov models for time-specific clustering, with an analysis of the well being of nations. J. R. Stat. Soc. C 68, 603–621.

Bartolucci, F., Farcomeni, A., 2009. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. J. Amer. Statist. Assoc. 104, 816–831.

Bartolucci, F., Farcomeni, A., 2015. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics 71, 80–89.

Bartolucci, F., Farcomeni, A., 2021. A hidden Markov space-time model for mapping of the dynamics of global access to food.

Bartolucci, F., Farcomeni, A., Pennoni, F., 2013. Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC Press, Boca Raton, FL.

Bartolucci, F., Farcomeni, A., Pennoni, F., 2014. Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. TEST 23, 433–465.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 192–225.

Besag, J., 1975. Statistical analysis of non-lattice data. J. R. Stat. Soc. D 24, 179–195.

Buss, L., Prete, C., Abrahim, C., Mendrone, A., Salomon, T., de Almeida-Neto, C., Franca, R., Belotti, M., Carvalho, M., Costa, A., Crispim, M., Ferreira, S., Fraiji, N., Gurzenda, S., Whittaker, C., Kamaura, L., Takecian, P., da Silva Peixoto, P., Oikawa, M., Nishiya, A., Rocha, V., Salles, N., de Souza Santos, A., da Silva, M.A., Custer, B., Parag, K., Barral-Netto, M., Kraemer, M., Pereira, R., Pybus, O., Busch, M., Castro, M., Dye, C., Nascimento, V., Faria, N., Sabino, E., 2021. Three-quarters attack rate of SARS-CoV-2 in the Brazilian amazon during a largely unmitigated epidemic. Science 371, 288–292.

Cabras, S., 2020. A Bayesian deep learning model for estimating COVID-19 evolution in Spain. arXiv:2005.10335.

Contreras, S., Dehning, J., Loidolt, M., Zierenberg, J., Spitzner, F.P., Urrea-Quintero, J.H., Mohr, S.B., Wilczek, M., Wibral, M., Priesemann, V., 2021. The challenges of containing SARS-CoV-2 via test-trace-and-isolate. Nature Commun. 12, 378.

Del Sole, F., Farcomeni, A., Loffredo, L., Carnevale, R., Menichelli, D., Vicario, T., Pignatelli, P., Pastori, D., 2020. Features of severe COVID-19: a systematic review and meta-analysis. Eur. J. Clin. Investig. 50, e13378.

Della Rossa, F., Salzano, D., Di Meglio, A., De Lellis, F., Coraggio, M., Calabrese, C., Guarino, A., Cardona-Rivera, R., De Lellis, P., Liuzza, D., Lo Iudice, F., Russo, G., di Bernardo, M., 2020. A network model of Italy shows that intermittent regional strategies can alleviate the COVID-19 epidemic. Nature Commun. 11, 5106.

Dotto, F., Farcomeni, A., Pittau, M.G., Zelli, R., 2019. A dynamic inhomogeneous latent state model for measuring material deprivation. J. R. Stat. Soc. A 182, 495–516.

Everitt, R.G., 2012. Bayesian parameter estimation for latent Markov random fields and social networks. J. Comput. Graph. Statist. 21, 940–960.

Farcomeni, A., 2015. Generalized linear mixed models based on latent Markov heterogeneity structures. Scand. J. Stat. 42, 1127–1135.

Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G., Lovison, G., 2021a. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom. J. 63, 503–513.

Farcomeni, A., Ranalli, M., Viviani, S., 2021b. Dimension reduction for longitudinal multivariate data by optimizing class separation of projected latent Markov models. Test in press.
Friel, N., Pettitt, A., 2004. Likelihood estimation and inference for the autologistic model. J. Comput. Graph. Statist. 13, 232–246.
Friel, N., Pettitt, A., Reeves, R., Wit, E., 2009. Bayesian inference in hidden Markov random fields for binary data defined on large lattices. J. Comput. Graph. Statist. 18, 243–261.
Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., Ventura, L., 2020. Robust inference from robust tsallis score: application to COVID-19 contagion in Italy. STAT.
Grasselli, G., Pesenti, A., Cecconi, M., 2020. Critical care utilization for the COVID-19 outbreak in lombardy, Italy: Early experience and forecast during an emergency response. JAMA 323, 1545–1546.
Green, P.J., Richardson, S., 2002. Hidden Markov models and disease mapping. J. Amer. Statist. Assoc. 97, 1055–1070.
Hu, B., Guo, H., Zhou, P., Shi, Z.-L., 2020. Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. in press.
Langrock, R., Kneib, T., Glennie, R., Michelot, T., 2017. Markov-switching generalized additive models. Stat. Comput. 27, 259–270.
Langrock, R., Kneib, T., Sohn, A., DeRuiter, S.L., 2015. Nonparametric inference in hidden Markov models using P-splines. Biometrics 71, 520–528.
Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J., 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 368, 489–493.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, CRC, London.
Qian, W., Titterington, D., 1991. Estimation of parameters in hidden Markov models. Philos. Trans. R. Soc. Lond. A 337, 407–428.
Spezia, L., Brewer, M.J., Birkel, C., 2017. An anisotropic and inhomogeneous hidden Markov model for the classification of water quality spatio-temporal series on a national scale: The case of Scotland. Environmetrics 28, e2427.
Spezia, L., Friel, N., Gimona, A., 2018. Spatial hidden Markov models and species distributions. J. Appl. Stat. 45, 1595–1615.
Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Amer. Statist. Assoc. 82, 528–540.
Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432.
Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.
Wood, S.N., 2017. Generalized Additive Models: An Introduction With R. Chapman & Hall/CRC, London, U.K..