POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian modeling and clustering for spatio-temporal areal data: An application to Italian Fertility Rate

Bayesian Statistics Project for
Mathematical Engineering

Federico Busetto, Daniele Fedrizzi, Niccolò Fontana, Emre Kayalioglu, Riccardo Lazzarini, Anna Lunghi

Academic year:
2022-2023

Abstract: During the last years, Italy has been experiencing a significant decrease in the birth rate. To understand which socio-economic factors influence the most this trend, we have collected spatio-temporal data from the Istat site and analysed it using a Bayesian model. In particular, we made use of spatio-temporal random effects with a conditional autoregressive prior, where the temporal correlation is modeled through an autoregressive mean decomposition and the spatial correlation by the precision matrix inheriting the neighboring structure of the data. Moreover we clustered the areal locations by means of a Bayesian nonparametric approach. In particular, we have modelled the vectors of areal specific coefficients of the regressors as being distributed as a Dirichlet Process, allowing us to cluster together the areal locations with same coefficients. The implementation of the model was performed entirely on `STAN`, while we used the library `salso` on R to perform posterior inference on the cluster allocation vector. We eventually compared, through predictive goodness of fit criteria our proposed model to a more classical CAR model implemented in the R library `CARBayesST`.

Key-words: Bayesian Nonparametrics, Conditional Autoregressive Priors, Markov Chain Monte Carlo, Spatio-Temporal Areal Unit Modeling, STAN

1. Italian Birth Rate

In 2021 Italy has reached an all time low in the number of births: for the first time ever less than 400.000 children were born during the year, the lowest value since 1861. This trend is certainly not a surprise; in fact, it is recorded that in the span of 10 years, between 2008 and 2019, the number of births dropped by 27%. The reasons for this decline are complex and multifaceted, but some factors that have been identified include economic instability, increased access to education and career opportunities for women, and changing social attitudes toward family and childbearing.
In a survey conducted by OpenPolis [5], it was recorded that women under 30 are the group that experiences economic insecurity the most and that 66% of them feel that their life plans are put at risk. Moreover women in Italy, as well as in other developed countries, are increasingly pursuing higher education and career opportunities and choosing to delay having children until later in life, or choosing not to have children at all. Additionally, gender inequalities in the labor market also play a role. Women are often paid less than men for the same work, and they are underrepresented in leadership positions. This can make it difficult for women to achieve financial

stability, which can further discourage them from having children.
There are also regional differences in birth rates within Italy. For example, the southern regions of Italy tend to have lower birth rates than the northern regions. This can be attributed to a number of factors, including economic disparities between the north and south, as well as cultural and social differences. For instance, Lombardia and Emilia-Romagna have higher birth rates than the national average, thanks to a more favorable economic situation and better job opportunities, while the province of Caserta went from having 14.95 children born per thousand inhabitants in 1990 to 7.9 in 2020 [2].

1.1. ISTAT Dataset

ISTAT gives access through `www.dati.istat.it` to thousands of data. We decided, in order to explain with socio-economic indices the phenomenon of declining birth rates, to enrich the data on fertility in Italy with data concerning the age, inactivity and employment of women and men. In particular we have data for $I = 106$ italian provinces and $T = 10$ years from 2011 to 2020. For each province $i = 1, \dots, I$ and each year $t = 1, \dots, T$ we have:

- `tasso.di.fecondita.totale`: this is the variable of interest. Fertility rate represents the average number of children per woman of childbearing age (15-49 years);
- `eta.media.delle.madri.al.parto`: average age of the mother at delivery;
- `eta.media.dei.padri.alla.nascita.del.figlio`: average age of the father at birth of the child;
- `tasso.di.inattivita.delle.femmine`: The inactivity rate is an index concerning dependents and people of working age. It is calculated as the division between the population under 16 and over 64, and the working-age population (over 15 and under 64). In this case it concerns only females;
- `tasso.di.inattivita.dei.maschi`: inactivity rate for males only;
- `tasso.di.inattivita.totale`: global inactivity rate of the province;
- `tasso.di.occupazione.delle.femmine`: the employment rate measures the incidence of employed people in the total population. It is obtained from the ratio of those employed between the ages of 15 and 64 to the population of the same age group. In this case it concerns only females;
- `tasso.di.occupazione.dei.maschi`: employment rate for males only,
- `tasso.di.occupazione.totale`: global employment rate of the province.

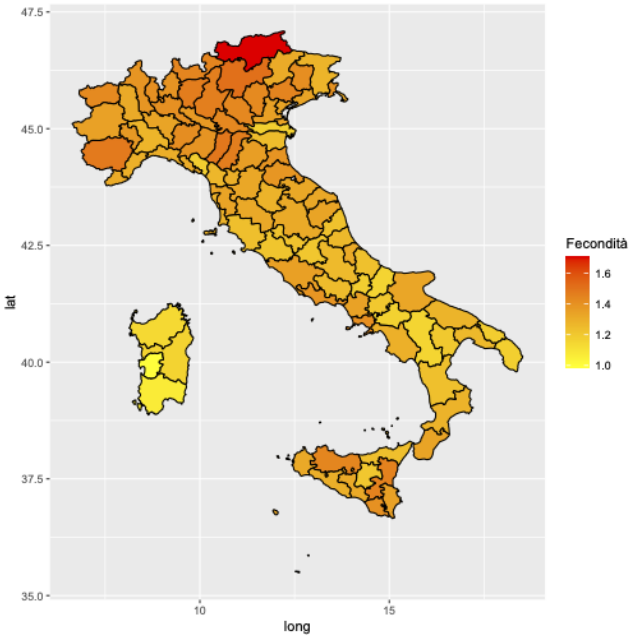


Figure 1: Average magnitude of the fertility rate across Italian provinces

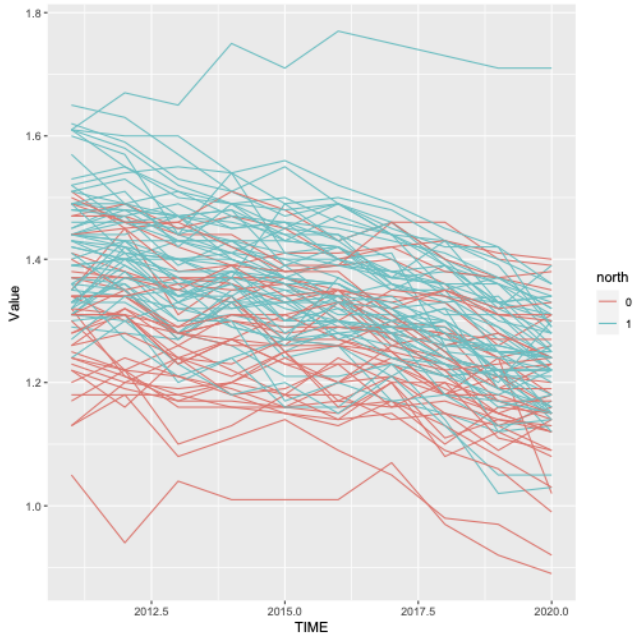


Figure 2: Fertility rate over time from 2011 to 2020 for each province divided by north and south

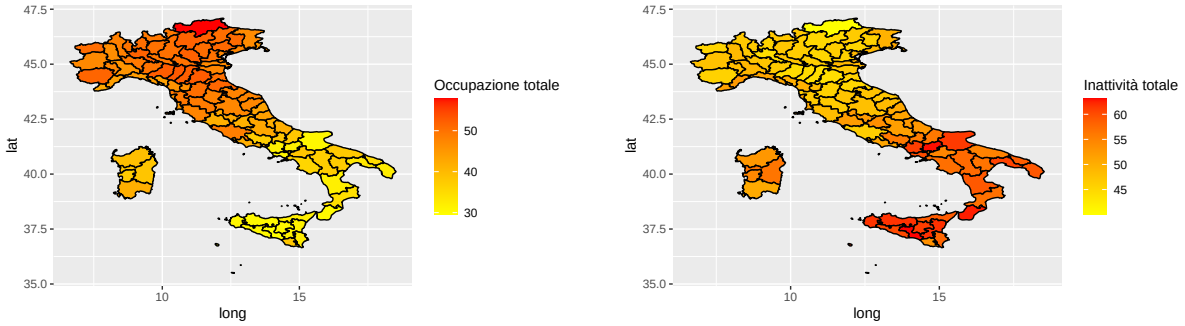
Figure 1 and figure 2 show the difference in provinces and time of the fertility rate. More precisely it is clear from figure 2 that the trend of the rate is decreasing over time and that on average the north has greater values than the south. The latter is also highlighted by figure 1 as we can see the north is associated to higher fertility rates.
It is interesting to note that from figure 2 it appears that a province in the north is not affected at all by the birth rate decrease. This province is the province of Bolzano, which is a peculiar region of Italy since it is autonomous province and it is mostly inhabited by native German speakers. Bolzano happens also to be the province associated to the highest value of employment rate and lowest value of inactivity rate.

Table 1

	Mean	St.dev	Min	1st Qu.	3rd Qu.	Max
fertility rate	1.319	0.1230447	0.89	1.24	1.4	1.77
age of the mother	31.7	0.5481536	30.13	31.34	32.08	33.4
age of the father	35.37	0.5206872	33.72	35.09	35.69	37.42
female inactivity	60.17	7.504205	45.86	54.38	66.25	80.14
male inactivity	41.56	4.397646	31.8	38.18	44.68	56.68
total inactivity	51.2	5.775728	39.37	46.58	55.52	66.21
female employment	35.09	8.409269	15.74	27.88	41.52	52.45
male employment	52.72	6.51996	25.42	47.78	57.87	65.98
total employment	43.6	7.331285	25.69	37.64	49.52	58.87

Table 1: Summary statistics of the response and predictors

The dichotomy between north and south in Italy is better expressed by means of the economic variables employment rate and inactivity rate. In figure 3 we can see that the northern regions are the one associate to higher values of employment rate and lower values of inactivity rate, while the vice versa holds true for the southern regions. This results are not unexpected since there are significant economic differences between northern and southern Italy. For instance, the northern regions of Italy are generally more industrialized than the southern regions which are traditionally more agriculturally based.



(a) Total employment rate

(b) Total inactivity rate

Figure 3: Average magnitude of employment and inactivity rate across Italian provinces

1.2. Moran's I and Geary's C

To conclude an exploratory analysis on the dataset it is useful to investigate the correlations between variables. Figure 4 shows scatter plots, histograms, and Pearson correlations of the fertility rate and the other covariates, grouped into northern and southern regions. We can see that the north-south duality is still reflected on the economic variables while being less significant in the fertility rate and the age of mother and father.
Employment and inactivity rate are strongly negatively correlated as expected. Fertility rate is negatively correlated with both the age of the mother and the age of the father, suggesting a tendency for families to have younger parents. Fertility rate is also positively correlated with employment and negatively with inactivity rate; this can be explained as a need of financial stability for a couple to have children.

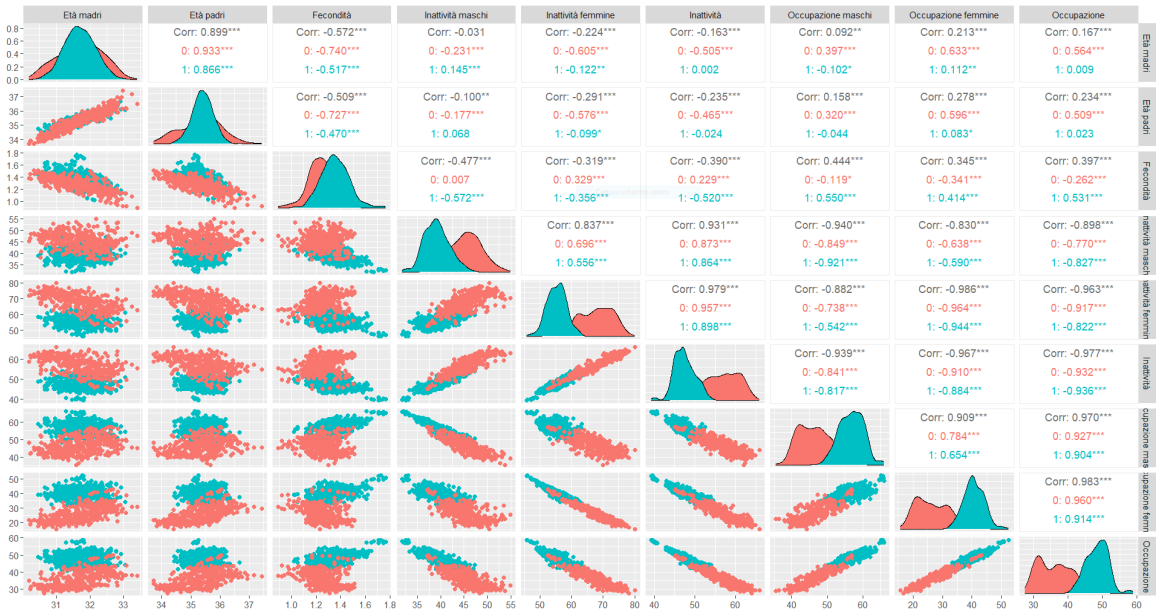


Figure 4: GGpairs of the dataset divided between north and south

Moran's I and Geary's C are indices that measure spatial autocorrelation. To compute them we need the proximity matrix W. W is an $I \times I$ binary matrix that describes the spatial structure of the areal locations. More precisely $W_{ij} = 1$ if province i and j are neighbours and $W_{ij} = 0$ if they are not. Moran's I and Geary's C are computed as follows:

$$\text{Moran's I} = \frac{n \sum_i \sum_j W_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} W_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

$$\text{Geary's C} = \frac{(n-1) \sum_i \sum_j W_{ij}(Y_i - \bar{Y})^2}{2(\sum_{i \neq j} W_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

where, for a given time $t = 1, \dots, T$, Y_i is the value in a certain site to which the indices refer, $\bar{Y}$ is the mean value across all areal locations and n the numerosity of the sample. Then we averaged the indices over time to obtain the following results:

Table 2

	Moran's I	**Geary's C**
fertility rate	0.4378463	0.4286952
age of the mother	0.3901207	0.5001999
age of the father	0.3519166	0.5467671
female inactivity	0.8496539	0.1227897
male inactivity	0.6619759	0.2628052
total inactivity	0.8217314	0.1440086
female employment	0.8588112	0.1089427
male employment	0.8010590	0.1487005
total employment	0.8564514	0.1087377

Table 2: Moran's I and Geary's C for spatial autocorrelation

Moran's I values from -1 to 0, indicate negative spatial autocorrelation, while values from 0 to 1 indicate positive spatial autocorrelation. Values of Geary's C between 0 and 1 indicate positive spatial autocorrelation, and values above 1 negative spatial autocorrelation. We can see that every variable has positive spatial correlation and the highest values of the indices are associated to economic features highlighting again the duality of north and south.

2. Model Specification

Following the model proposed by [4], in this section we introduce a bayesian hierarchical model able to examine spatio-temporal dependencies as well as uncover the grouping pattern in areal data. The former is obtained by including random effects through a CAR prior, the latter by using a Bayesian nonparametric approach.

2.1. Spatio-Temporal Random Effects

Let Y_{it} be the observed value of interest at the areal location $i = 1, \dots, I$ and at time $t = 1, \dots, T$. so that the response data can be rewritten in form of a $I \times T$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_T)$ where $\mathbf{Y}_t = (Y_{1t}, \dots, Y_{It})^\intercal$. In our case Y_{it} will be the fertility rate in province i at year t. Let also $\mathbf{x}_{it}$ be a $\mathbb{R}^{p+1}$ vector of p known covariates for areal unit i at time t, with first component equal to 1 to include the intercept. The model for Y_{it} is the following:

$$Y_{it} = \mathbf{x}_{it}^\intercal \boldsymbol{\beta}_i + w_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

where $\boldsymbol{\beta}_i$ is the vector of regression coefficients for the areal unit i, w_{it} is a spatio-temporal random effect and ε_{it} is a gaussian error with variance σ^2.
For the modeling of the spatio-temporal random effects we follow [1] where they represented the spatio-temporal structure with a multivariate first order autoregressive process with a spatially autocorrelated precision matrix. The model specification is as follows:

$$\begin{aligned} &\mathbf{w}_t | \mathbf{w}_{t-1} \sim N_I(\xi \mathbf{w}_{t-1}, \tau^2 Q(\rho, W)^{-1}) \\ &\mathbf{w}_0 \sim N_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1}) \end{aligned} \tag{2}$$

where $\mathbf{w}_t$ is a $\mathbb{R}^I$ vector consisting of the random effects at each areal location at time t, ξ is the autoregressive coefficient and the covariance matrix is composed of a scale parameter τ^2 and the inverse of $Q(\rho, W)$ which is a matrix modeling the spatial correlation.

For the choice of $Q(\rho, W)$ we followed [4] and took

$$Q(\rho, W) = \rho(\text{diag}(W\mathbf{1}) - W) + (1-\rho)\mathbb{I}_I$$

with W the proximity matrix previously described in 1.2, $\mathbf{1}$ a T-dimensional vector of ones and $\mathbb{I}_I$ the identity matrix of size $I \times I$. The matrix $\text{diag}(W\mathbf{1}) - W$ has the number of neighbours of each areal location on the diagonal and -1 on the element i, j if site i and site j are neighbours. The parameter $\rho \in [0, 1]$ models the correlation between random effects, in particular the lower the ρ is the more the random effect will be independent of each other, with the trivial case $\rho = 0$ being totally independent spatial random effect. As noted in [4] the interpretation of ρ sometimes can be difficult so in the following we assume that $\rho = 0.95$ to enforce spatial dependence among random effects. As a consequence of the choiche of $Q(\rho, W)$, the joint law of the vector of random effects will be a Gaussian Markov Random Field (GMRF) with covariance matrix reflecting the spatial disposition of the areal locations and symmetric positive definite thanks to the addition of ρ.

2.2. Bayesian Nonparametric Clustering

To perform clustering we include a Dirichlet Process prior for some areal-specific parameters.
Thanks to the Stick-breaking construction we can say that a random probablity measure P follows a Dirichlet process of parameters $\alpha > 0$ and P_0 if:

$$P = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}$$

where δ_{θ_j} is the dirac delta measure centered in θ_j and ω_j are the weights given by the stick-breaking construction. More precisely:

$$\begin{aligned}
&\theta_j \overset{iid}{\sim} P_0 \\
&\omega_j = V_j \prod_{k=1}^{j-1}(1 - V_k) \quad j = 2, 3, \dots \\
&\omega_1 = V_1 \\
&V_k \overset{iid}{\sim} \text{Beta}(1, \alpha)
\end{aligned}$$

The interpretation of the parameters of the dirichlet process is simple, indeed we can say P_0 is the mean measure of P and α measures the variability around P_0.
Thanks to the discrete nature of the random probability measure $P \sim DP(\alpha, P_0)$ we allow for clustering in a straightforward way, indeed by placing a dirichlet process prior on the $\boldsymbol{\beta}_i$ we have:

$$\begin{aligned}
&\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_I | P \overset{iid}{\sim} P \\
&P \sim DP(\alpha, P_0)
\end{aligned}$$

that can be rewritten introducing the cluster allocation vector $\mathbf{s} = (s_1, \dots, s_I)$ and a set of unique values $(\boldsymbol{\beta}_1^\star, \dots, \boldsymbol{\beta}_{K_I}^\star)$ with $K_I < I$ by using the following "clustering rule":

$$s_i = j \iff \boldsymbol{\beta}_i = \boldsymbol{\beta}_j^\star$$

By doing so we can recognize at most K_I clusters each of which has its own distinct value of the regression parameters $\boldsymbol{\beta}_j^\star$. The cluster allocation vector $\mathbf{s}$ a priori will follow a Polya Urn scheme, that is:

$$\begin{aligned}
p(\mathbf{s}|\alpha) &= p(s_1) \prod_{i=2}^{I} p(s_i | s_1, \dots, s_{i-1}) \\
p(s_i | s_1, \dots, s_{i-1}) &= \begin{cases} \frac{n_{ij}}{i-1+\alpha} & j = 1, \dots, K_I \\ \frac{\alpha}{i-1+\alpha} & j = K_I + 1 \end{cases}
\end{aligned}$$

where n_{ij} is the size of the jth cluster before we assign the ith observation, that is the size of $\{l < i | s_l = j\}$. This means that the cluster of a given areal location i will be either one of the previously observed cluster or a new additional cluster, with the first areal location always being put in cluster number 1.

2.3. Full Model

The model described before is then completed by choosing prior distributions for the remaining parameters:

$$
\begin{aligned}
&Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}^{\star}_{s_i}, w_{it}, \sigma^2, s_i \overset{\text{ind}}{\sim} N(\mathbf{x}_{it}^{\top}\boldsymbol{\beta}^{\star}_{s_i} + w_{it}, \sigma^2) \\
&\mathbf{w}_t|\mathbf{w}_{t-1}, \xi, \tau^2, \rho, W \sim N_I(\xi\mathbf{w}_{t-1}, \tau^2 Q(\rho, W)^{-1}) \\
&\mathbf{w}_1|\tau^2, \rho, W \sim N_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1}) \\
&\xi \sim \text{Beta}_{(-1,1)}(a_\xi, b_\xi) \\
&\sigma^2 \sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
&\tau^2 \sim \text{Inv-Gamma}(a_{\tau^2}, b_{\tau^2}) \\
&\mathbf{s}|\alpha \sim \text{PòlyaUrn}(\mathbf{s}|\alpha) \\
&\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \\
&\boldsymbol{\beta}^{\star}_1, \ldots, \boldsymbol{\beta}^{\star}_{K_I}|\boldsymbol{\mu}_\beta, \Sigma_\beta, \overset{\text{iid}}{\sim} P_0, \\
&P_0(d\boldsymbol{\beta}^{\star}) = N_{p+1}(d\boldsymbol{\beta}^{\star}|\boldsymbol{\mu}_0, \Sigma_0)
\end{aligned}
\tag{3}
$$

note that $\xi \sim \text{Beta}_{(-1,1)}(a_\xi, b_\xi)$ is the Beta distribution over the interval $(-1, 1)$ in other words $\xi = 2\eta - 1$ with $\eta \sim \text{Beta}(a_\xi, b_\xi)$.

3. Application to Italian Fertility Rate

The aim of this section is to analyse the Italian fertility dataset in 1.1 using model (3). In section 3.1 we discuss some useful methods to fix the hyperparameters of the prior distributions in order to obtain insightful results about the clustering of Italian provinces. Eventually, we describe one of the main model competitors for spatio-temporal areal unit modeling, that is the one obtained through the function `ST.CARar()` in the R-package `CARBayesST` and make some comparison with our proposed model.

3.1. Prior Elicitation

We chose to set the hyperparameters of τ^2 and σ^2 to $a_{\tau^2} = 3$, $b_{\tau^2} = 2$ and $a_{\sigma^2} = 3$, $b_{\sigma^2} = 2$ implying for both prior expected value and prior variance equal to 1, as in [4].
After an augmented Dickey-Fuller test on response variable for each province, we can conclude that the fertility rate is not a stationary time-series, we decided then not to use a prior with bounded support for ξ, as a $\text{Beta}_{(-1,1)}(a_\xi, b_\xi)$, but rather a normal distribution $N(\mu_\xi, \sigma^2_\xi)$ centered around 0 and with variance equal to 4 ($\mu_\xi = 0$, $\sigma^2_\xi = 4$).
To help the model in recognizing clusters, it is useful to put informative priors on the clustered parameters, $\boldsymbol{\beta}^{\star}_j$ in our case. To achieve so we decided to put a prior on the parameters, $\boldsymbol{\mu}_0$ and Σ_0, of the base measure P_0 of the Dirichlet process. In particular:

$$
\begin{aligned}
&P_0(d\boldsymbol{\beta}^{\star}) = N_{p+1}(d\boldsymbol{\beta}^{\star}|\boldsymbol{\mu}_0, \Sigma_0) \\
&\boldsymbol{\mu}_0 \sim N_{p+1}(\boldsymbol{\eta}_0, \tau_0^2 \mathbb{I}_{p+1}) \\
&\Sigma_0 = \sigma_0^2 \mathbb{I}_{p+1} \\
&\sigma_0^2 \sim \text{Inv-Gamma}(a_{\sigma_0^2}, b_{\sigma_0^2})
\end{aligned}
\tag{4}
$$

To tune the hyperparameters of these new set priors, we fitted a standard OLS regression model for each province, giving us I vectors of regression coefficients, $\boldsymbol{\beta}^{\text{OLS}}_1, \ldots, \boldsymbol{\beta}^{\text{OLS}}_I$. Starting from these we put $\boldsymbol{\eta}_0$ equal to the sample mean of the $\boldsymbol{\beta}^{\text{OLS}}_j$ and τ_0^2 large enough to contain all the coefficients. While we fixed $a_{\sigma_0^2}$ and $b_{\sigma_0^2}$ such that the prior mean of σ_0^2 was equal to the mean of the diagonal of the sample covariance matrix of the $\boldsymbol{\beta}^{\text{OLS}}_j$ and with moderate prior variance.
Following [4], we fixed the hyperparameters of α to $a_\alpha = 3$ and $b_\alpha = 2$ in order to obtain a prior expected number of clusters of 6.75 with prior variance of number of clusters equal to 7.32. These values not only help the model in finding clusters but are also reflecting the division of Italian provinces with respect to the economic variables.

3.2. Optimal Partition of Provinces

In this section we briefly describe one of the main methods to obtain cluster estimates from draws from the posterior distribution of the cluster allocation vector $\mathbf{s}$. Following the idea of the posterior mean as the minimizer of a quadratic loss function, we can define the cluster estimate as a clever summary statistics of the posterior distribution, that is the minimizer of an ad hoc loss function. More precisely:

$$\hat{\mathbf{s}} = \arg\min_{\tilde{\mathbf{s}}} \mathbb{E}_{\mathbf{s}}[L(\mathbf{s}, \tilde{\mathbf{s}})|\text{Data}]$$

In the literature we find many choices for the loss function, two of the most popular ones are the Binder's loss and the Variation of Information (VI). These two losses have the important property of being invariant with respect to label-switching, giving the opportunity of working directly on the cluster allocation vectors.
The definition of the Binder's loss is the following:

$$L_{\text{Binder}}(\mathbf{s}, \tilde{\mathbf{s}}) = \sum_{i,j=1}^{n} \Big[a \cdot I(s_i = s_j) I(\tilde{s}_i \neq \tilde{s}_j) + b \cdot I(s_i \neq s_j) I(\tilde{s}_i = \tilde{s}_j) \Big]$$

where a and b are hyperparameters to tune. a is the cost for two areal locations classified in different clusters when they actually belong to the same cluster, while b is the cost for two areal locations classified in the same cluster when they actually belong to the different clusters. Indeed, if $a \gg b$ then the Binder's loss will estimate very few clusters as we favor the cost of misclassifying observations in the same cluster, on the contrary if $a \ll b$ the Binder's loss will estimate more clusters as we favor the cost of misclassifying observations in different clusters. In practice, often a and b are taken equal. This choice also facilitate the task of minimization since we have a way to rewrite the expected value of the Binder's loss in a easier way:

$$\mathbb{E}\Big[L_{\text{Binder}}(\mathbf{s}, \tilde{\mathbf{s}})|\text{Data} \Big] = \sum_{i,j=1}^{n} I(\tilde{s}_i \neq \tilde{s}_j) \cdot \left(p_{ij} - \frac{1}{2} \right)$$

$$p_{ij} = \mathbb{P}(s_i = s_j|\text{Data}) \approx \frac{1}{M} \sum_{m=1}^{M} I(s_i^{(m)} = s_j^{(m)})$$

VI loss is based on information theory and measures the amount of information that is lost when going from the true class labels to the predicted class labels. It is defined as the sum of the entropy of the true class labels, the entropy of the predicted class labels and their joint entropy suitably weighted:

$$L_{\text{VI}}(C, \tilde{C}) = a \sum_{C \in \mathcal{C}} \frac{|C|}{n} \log_2 \left(\frac{|C|}{n} \right) + b \sum_{\tilde{C} \in \tilde{\mathcal{C}}} \frac{|\tilde{C}|}{n} \log_2 \left(\frac{|\tilde{C}|}{n} \right) - (a+b) \sum_{C \in \mathcal{C}} \sum_{\tilde{C} \in \tilde{\mathcal{C}}} \frac{|C \cap \tilde{C}|}{n} \log_2 \left(\frac{|C \cap \tilde{C}|}{n} \right)$$

where C and $\tilde{C}$ are, respectively, the true and estimated partitions and |C|, $|\tilde{C}|$ their sizes.

To perform the minimization task we relied on the R-package `salso` that has implemented both the Binder and the VI losses and uses a stochastic search algorithm to find the optimum.

3.3. Posterior Inference

To analyse the results of the application to the Italian fertility rate we run four MCMCs on the standardized variables with 10000 burn-in iterations and we keep the following 10000 iterations. After having found the optimal partition with VI loss, we re-run the MCMCs conditionally on the VI loss partition with same number of iterations of the first run.
Figure 5 shows the clustering found on the Italian provinces, while table 3 summarizes the posterior means of the regression coefficients for each cluster.
Despite cluster 1 being more concentrated toward the northern regions and cluster 2 toward the centre-southern regions, the found clusters seem not to be following the geographical positioning of the provinces but rather the economical and social conditions. Most of all, as figure 5 shows, the division between clusters favors the "quality" of the fertility rate. Indeed we can find three distinct groups of fertility rate:

- High: clusters for which the 25% quantile of fertility rate is greater than 1.35
- Medium: clusters for which the median of fertility rate is greater than 1.35
- Low: clusters for which the 75% quantile of fertility rate is less than 1.35

Cluster 1 belongs to the medium group, clusters 2 and 4 belong to the low group and clusters 3, 5 and 6 belong to the high group.

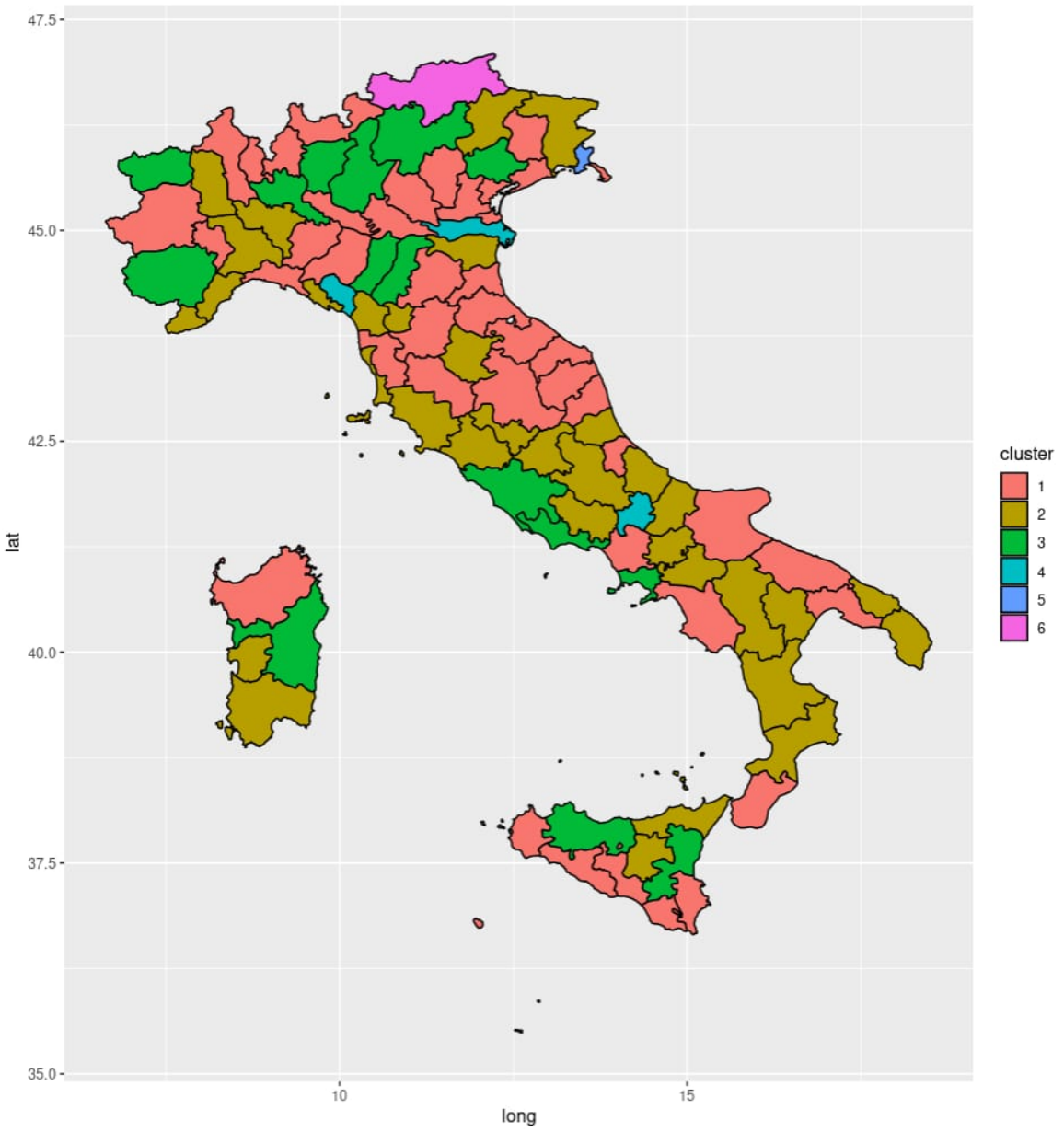
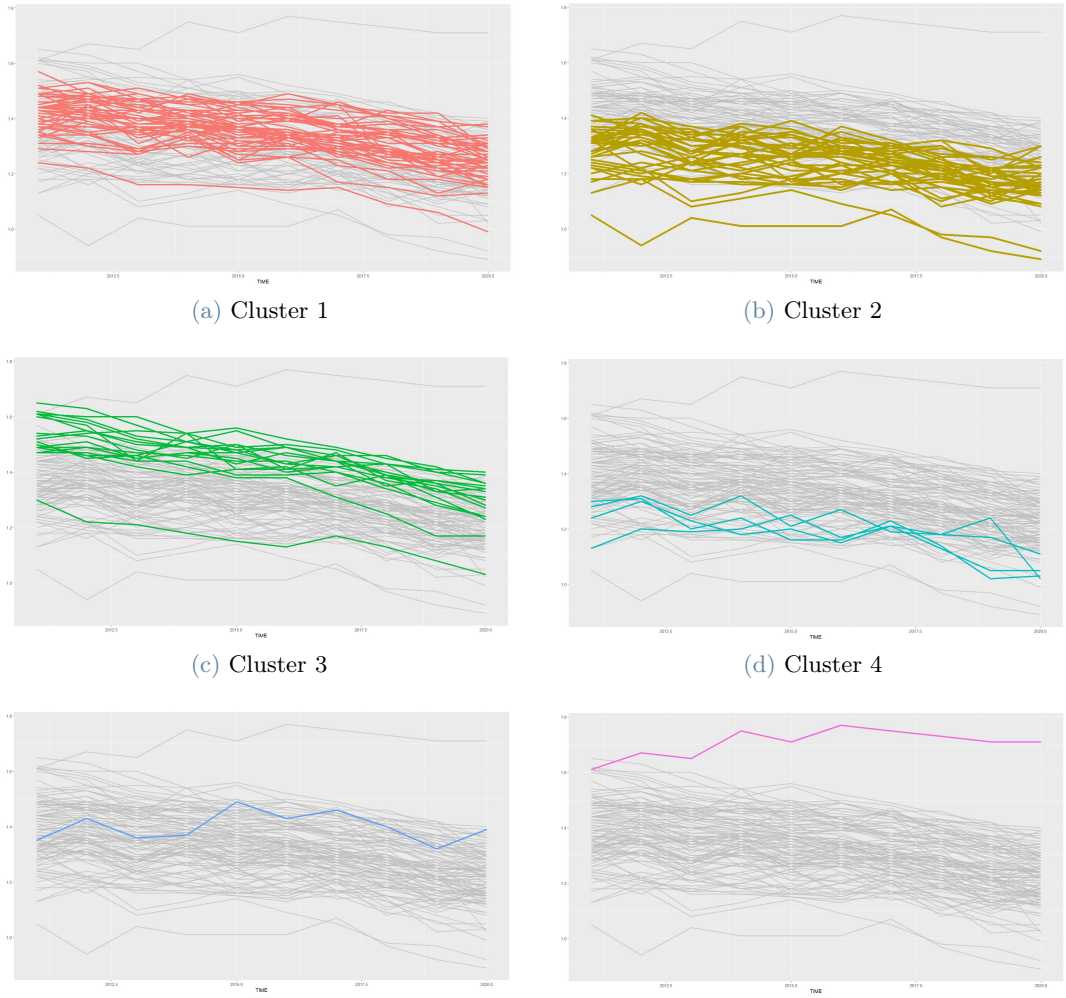


Figure 5: Partition of the Italian provinces obtained by minimizing the posterior expectation of the VI loss function

Table 3

(Size)	Intercept	Age Mother	Female Inac.	Male Inac.	Female Emp.	Male Emp.
1 (47)	0.447	-0.314	0.120	-0.129	0.259	0.021
2 (36)	-0.273	-0.256	0.097	0.089	0.331	0.102
3 (17)	1.038	-0.392	0.401	-0.151	0.424	0.094
4 (4)	-0.693	-0.184	-0.292	-0.500	-0.221	-0.601
5 (1)	0.599	0.016	0.061	-0.797	-0.077	-0.050
6 (1)	0.577	0.675	-0.252	-0.423	0.559	-0.086

Table 3: Posterior mean of the regression coefficients for each cluster for the standardized variables



(a) Cluster 1 (b) Cluster 2

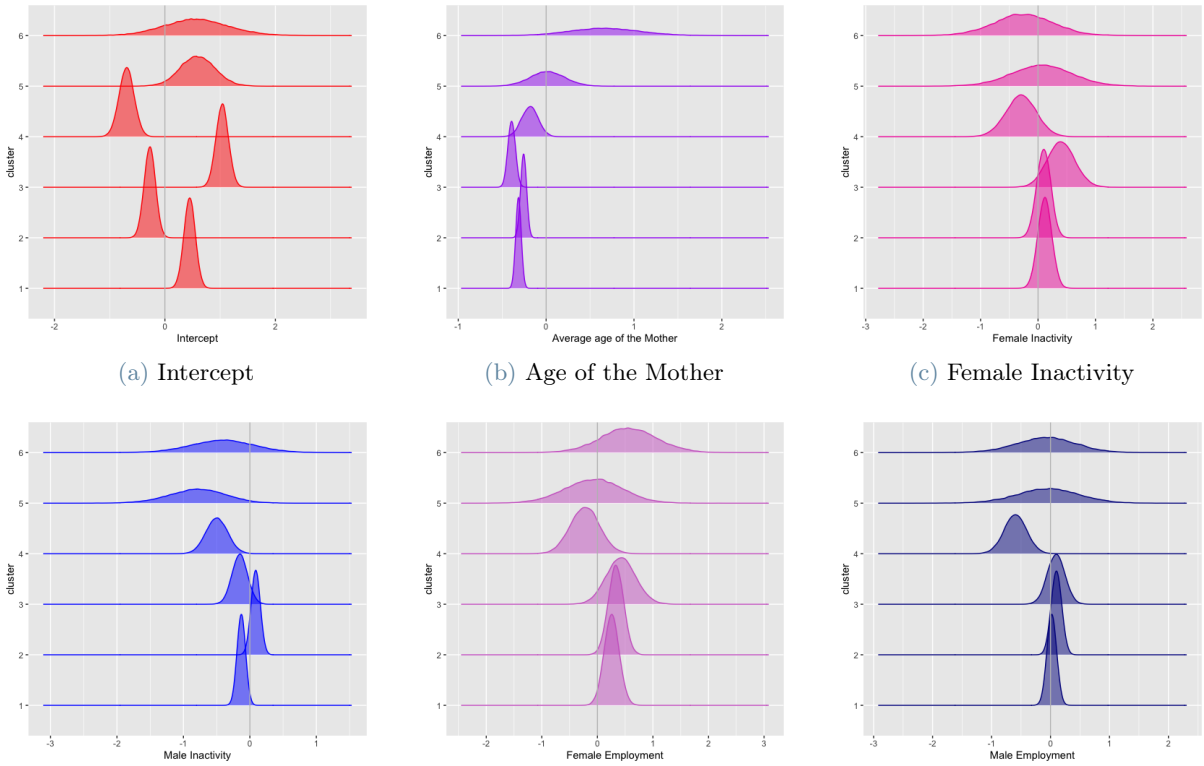
(c) Cluster 3 (d) Cluster 4

(e) Cluster 5 (f) Cluster 6

Figure 6: Fertility rate time-series divided for each cluster

The largest cluster, cluster 1, is mainly concentrated toward the centre-northern regions of Italy (Mantova, Parma and Torino belong to this cluster) but also includes southern provinces like Bari, Crotone, Caserta and Siracusa. In fact almost all the most southern provinces of Sicily are in cluster 1. Cluster 2 is mostly spread in the centre-southern regions of Italy, with provinces like L'Aquila, Matera and Vibo-Valentia belonging to this cluster, but it also has representative provinces in the north like Alessandria, Ferrara and Pavia. Cluster 3 includes the most important cities of Italy in terms of economic power like Milano, Roma and Napoli and other smaller cities with high GDP like Brescia and Monza. Cluster 4 is made of four provinces, Biella, Isernia, Massa-Carrara and Rovigo. Cluster 5 is a singleton containing the province of Gorizia. Cluster 6 is another singleton containing the province of Bolzano.
We can notice that, aside from clusters 5 and 6, there is an intuitive tendency of having lower values for the average age of the mothers associated to higher values of fertility rate, as women prefer to have children not too late in the childbearing age.
The unfavorable nature of the fertility rate for clusters 2 and 4 is confirmed through inspection of the intercepts, indeed in both clusters it is significantly negative as figure 7a shows. Nonetheless it is interesting to further investigate the distinction between these two clusters. Cluster 2 seems also to be associated to worse values for the economic variables, with strikingly high values of female and male inactivity. While cluster 4, despite having better economic features regarding employment and inactivity, is associated to the lowest values of fertility rate. Cluster 3 is the cluster with highest values for female and male employment, and it has a clear tendency of having higher values of fertility rate associated to higher values of the employment rate. This is confirmed by the fact that in cluster 3 we find the provinces with highest GDP of all Italy.
Cluster 1, despite not having much worse economic features of cluster 3, is associated to fertility rates that are not as high as in cluster 3. This is possibly explained by the fact the cluster 1 contains provinces with high employment but not as economically developed as the ones in cluster 3.
Cluster 5 is the province of Gorizia. This a surprising province, since it has good economic features, as high

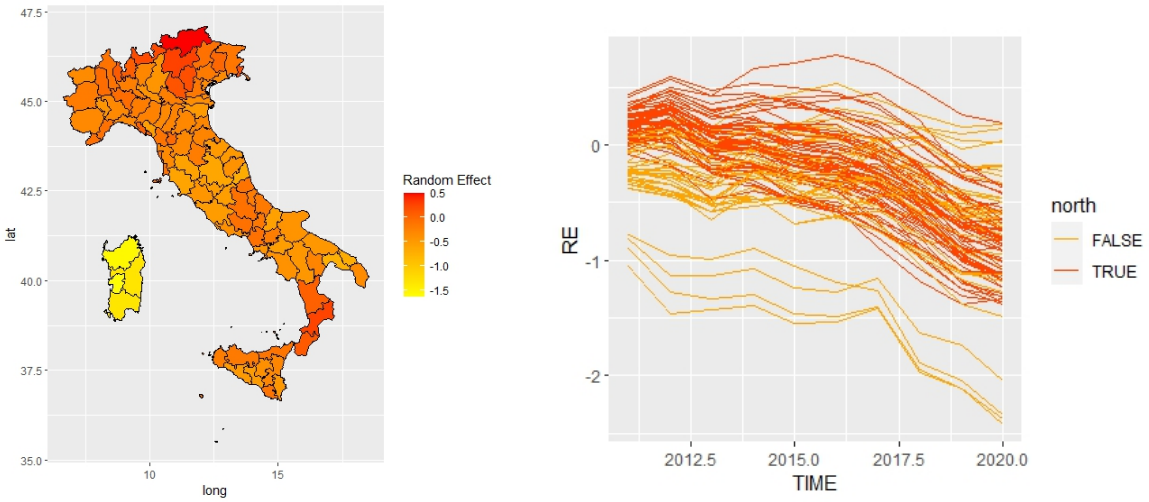
as in cluster 1, but it is the only province, aside from Bolzano, that has a positive trend in the fertility rate. Indeed the fertility rate of Gorizia in 2011 was of 1.35 and in 2020 was of 1.39, with the highest fertility rate touched in 2015 by the value of 1.49.
Cluster 6 is the autonomous province of Bolzano. As already noted in 1.1, Bolzano is a peculiar Italian province. It comes as no surprise that, since its fertility rate trend is positive and the highest among all other Italian provinces, it is associated to the best economic features, such as female and male employment. Moreover, Bolzano is the only province in which the regression coefficient related to the average age of the mother is significantly positive, this means that in the province women choose on average to have children later in the childbearing age. Bolzano, in this context, is really a stand alone province and the fact that its cluster is a singleton is a sign of the goodness of the model.



(a) Intercept (b) Age of the Mother (c) Female Inactivity

(d) Male Inactivity (e) Female Employment (f) Male Employment

Figure 7: Posterior kernel density estimates of the regression coefficients



(a) Average of spatio-temporal random effects for each province

(b) Time series of spatio-temporal random effects

Figure 8: Estimated spatio-temporal random effects

Figure 8 shows the estimated spatio-temporal random effects. As one can notice the division between north and south regions is still found as the northern regions seem to have higher fertility rate over time. The general tendency of the time series of the random effects is mostly a downwards trend, resembling the decrease in the fertility rate over the years. It is interesting, also, to notice that the provinces of the region of Sardegna have all the lowest values of the random effects, meaning that the spatio-temporal contribution to the fertility rate is strongly negative.

3.4. Competitor Models

In this section we describe one of the main competitor to the model in (3), that is the one obtained through the function `ST.CARar()` implemented in the R-package `CARBayesST` [3]. The competitor model has the following specification:

$$
\begin{aligned}
&Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}, w_{it}, \sigma^2 \overset{\text{ind}}{\sim} N(\mathbf{x}_{it}^\top \boldsymbol{\beta} + w_{it}, \sigma^2) \\
&\mathbf{w}_t|\mathbf{w}_{t-1}, \xi, \tau^2, \rho, W \sim N_I(\xi \mathbf{w}_{t-1}, \tau^2 Q(\rho, W)^{-1}) \\
&\mathbf{w}_1|\tau^2, \rho, W \sim N_I(\mathbf{0}, \tau^2 Q(\rho, W)^{-1}) \\
&\boldsymbol{\beta} \sim N_{p+1}(\mu_\beta, \Sigma_\beta) \\
&\xi \sim U([0,1]) \\
&\rho \sim U([0,1]) \\
&\sigma^2 \sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
&\tau^2 \sim \text{Inv-Gamma}(a_{\tau^2}, b_{\tau^2})
\end{aligned}
\tag{5}
$$

The `ST.CARar()` model follows closely model (3) but it lacks the Dirichlet Process prior that allows for clustering. Indeed model (5) uses a temporal autoregressive process of order one with a spatially autocorrelated precision matrix $Q(\rho, W) = \rho(\text{diag}(W\mathbf{1}) - W) + (1-\rho)\mathbb{I}_I$ for the random effects but employs a unique $\boldsymbol{\beta}$ of regression coefficients for all areal locations.
In order to assess the performance of model (3), we compute some predictive goodness of fit indices and compare it to the ones from `ST.CARar()`. In particular we computed the Log Pseudo-Marginal Likelihood (LPML) and Widely Applicable Information Criterion (WAIC).
In particular, the WAIC is obtained by calculating the log pointwise posterior predictive density (6) and then adjusting for overfitting with the correction term (7):

$$
\text{comp\_lppd} = \sum_{i=1}^{I} \log\left(\frac{1}{M}\sum_{m=1}^{M} p(\mathbf{Y}_i|\boldsymbol{\theta}^m)\right) \tag{6}
$$

$$
\text{comp\_p}_{\text{WAIC}} = 2\sum_{i=1}^{I}\left(\log\left(\frac{1}{M}\sum_{m=1}^{M} p(\mathbf{Y}_i|\boldsymbol{\theta}^m)\right) - \frac{1}{M}\sum_{m=1}^{M}\log p(\mathbf{Y}_i|\boldsymbol{\theta}^m)\right) \tag{7}
$$

$$
\text{WAIC} = -2\,(\text{comp\_lppd} - \text{comp\_p}_{\text{WAIC}}) \tag{8}
$$

where we denoted by $\boldsymbol{\theta}$ the vector of all unknown parameters and by $\boldsymbol{\theta}^m$ the mth posterior draw of $\boldsymbol{\theta}$ from a total of M MCMC draws.
While the definition of the LPML is as follows:

$$
\text{LPML} = \sum_{i=1}^{I} \log p(\mathbf{Y}_i|\mathbf{Y}_{(-i)})
$$

Given these definitions, the best model is the one minimizing WAIC and maximizing LPML. Table 4 summarises the computed indices.

Table 4

	WAIC	**LPML**
`ST.CARar()`	-29	-191
Model (3)	4	-62

Table 4: WAIC and LPML for model comparison

From the previous table we can conclude that there is no clear best model, indeed `ST.CARar()` has lower WAIC while model (3) has higher LPML. If on one hand model (3) does not allow for a prior on the parameter ρ enforcing spatial dependence among random effects by keeping it fixed to 0.95, on the other hand `ST.CARar()` does not allow for the clustering we have obtained from model (3). In conclusion we can say that from a purely interpretative point of view, we prefer model (3) since it allowed us to make all the insightful reasoning we have showed in section 3.3.

4. Code Implementation

Since the implementation of the MCMC algorithm for sampling from posterior distributions was performed in `STAN`, we switched to a model formulation that did not include a Pólya Urn scheme as a prior for $\boldsymbol{s}$. The Pólya Urn scheme is in fact a discrete distribution and therefore incompatible with the Hamiltonian Monte Carlo, that is the MCMC method used by `STAN` to sample from the posteriors.

We therefore started with the following equivalent formulation, for all $i = 1, .., I$ and $t = 1, ..., T$:

$$
\begin{aligned}
&\boldsymbol{Y}_i | X_i, \boldsymbol{w}_i, \sigma^2, P \overset{\text{ind}}{\sim} \int_{\mathbb{R}^{p+1}} N_T(\boldsymbol{Y}_i |\ X_i\boldsymbol{\beta} + \boldsymbol{w}_i, \sigma^2 \mathbb{I}_{p+1}) P(d\boldsymbol{\beta}) \\
&P \sim DP(\alpha, P_0) \\
&\boldsymbol{w}_t | \boldsymbol{w}_{t-1}, \xi, \tau^2, \rho, W \sim N_I(\xi \boldsymbol{w}_{t-1}, \tau^2 Q(\rho, W)^{-1}) \\
&\boldsymbol{w}_1 | \tau^2, \rho, W \sim N_I(\boldsymbol{0}, \tau^2 Q(\rho, W)^{-1}) \\
&\xi \sim Beta_{(-1,1)}(a_\xi, b_\xi) \\
&\sigma^2 \sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}) \\
&\tau^2 \sim \text{Inv-Gamma}(a_{\tau^2}, b_{\tau^2}) \\
&\rho \sim \text{Beta}(\alpha_\rho, \beta_\rho) \\
&\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \\
&P_0(d\boldsymbol{\beta}) = N_{p+1}(d\boldsymbol{\beta} | \boldsymbol{\mu}_0, \Sigma_0)
\end{aligned}
\tag{9}
$$

where $N_T(\boldsymbol{Y}_i |\ X_i\boldsymbol{\beta} + \boldsymbol{w}_i, \sigma^2 \mathbb{I}_{p+1})$ denotes the density of a multivariate Gaussian distribution of mean $X_i\boldsymbol{\beta} + \boldsymbol{w}_i$ and covariance matrix $\sigma^2 \mathbb{I}_{p+1}$, $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iT})^\top$, $\boldsymbol{w}_i = (w_{i1}, ..., w_{iT})^\top$ and X_i is a matrix of T rows, which are $\boldsymbol{x}_{it}^\top$ for $t = 1, .., T$.
Then, using the Stick-Breaking construction of the Dirichlet Process, the first row of (9) can be rewritten as:

$$
\boldsymbol{Y}_i | X_i, \boldsymbol{w}_i, \sigma^2 \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} \omega_j N_T(\boldsymbol{Y}_i |\ X_i\boldsymbol{\beta}_j + \boldsymbol{w}_i, \sigma^2 \mathbb{I}_{p+1})
$$

with ω_j as in (2.2) and $\boldsymbol{\beta}_j \overset{iid}{\sim} P_0$.
The above series can be suitably truncated to H in order to obtain a finite mixture model, where the mixturing elements are the coefficients of the linear regression. We chose H = 10 as a compromise between the goodness of the Dirichlet Process approximation and the performance of the code.
In this way we are assuming that the time series $\boldsymbol{Y}_i,\ i = 1, \ldots, I$ come from H different classes and that a priori

$$
\mathbb{P}(\boldsymbol{Y}_i \in \text{class } j) = \mathbb{P}(s_i = j) = \tilde{\omega}_j = \frac{\omega_j}{\sum_{j=1}^{H} \omega_j} \quad j = 1, \ldots, H
$$

Leading to the following posterior probabilities for the cluster allocation vector $\boldsymbol{s}$

$$
\mathbb{P}(s_i = j |\ X_i, \boldsymbol{Y}_i, \boldsymbol{w}_i, \sigma^2, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \tilde{\boldsymbol{\omega}}) = \frac{\tilde{\omega}_j N_T(\boldsymbol{Y}_i |\ X_i\boldsymbol{\beta}_j + \boldsymbol{w}_i, \sigma^2 \mathbb{I}_{p+1})}{\sum_{j=1}^{H} \tilde{\omega}_j N_T(\boldsymbol{Y}_i |\ X_i\boldsymbol{\beta}_j + \boldsymbol{w}_i, \sigma^2 \mathbb{I}_{p+1})}
$$

This allows us to obtain through the *generated quantities* of `STAN` draws from the posterior of $\boldsymbol{s}$ without including it directly in the model.

4.1. Code Optimization

Fixing $\rho = 0.95$ also implies fixing the covariance matrix of the random effects, which will therefore not change at each iteration of the MCMC.

It follows that we can factor, once and for all, the matrix $Q(\rho, W)^{-1}$ using the Cholesky decomposition, obtaining L such that $L^\top L = Q(\rho, W)^{-1}$ and we can reparametrize the random effects in the following more efficient way:

$$\begin{aligned} \boldsymbol{w}_1 &= L\boldsymbol{w}_1^{raw} \\ \boldsymbol{w}_t &= \xi \boldsymbol{w}_{t-1} + L\boldsymbol{w}_t^{raw} \\ \boldsymbol{w}_t^{raw} &\overset{\text{iid}}{\sim} N_I(\mathbf{0}, \tau^2 \mathbb{I}_{p+1}) \end{aligned} \tag{10}$$

Moreover the implementation of the model requires access to both rows and columns of the random effects matrix, which is why we decided to use a temporary matrix for $\boldsymbol{w}_{it}$ of type `matrix[I,T]` whose columns are accessible in an efficient manner and thus appropriate to the implementation of (10). Once this was done, since the construction of the target required access to $\boldsymbol{w}_i$ as a column vector, we transposed the matrix directly into a `matrix[T,I]` avoiding the transposed operator within the definition loops, that is more expensive in terms of memory and time.
Finally, we used *vectorization*, which is another technique provided by `STAN` to make the code more optimized (see [6]). This led us to choose appropriate data structures that could both make matrix operations possible and allow the use of factorized probability functions. For example, we declared the coefficients of the linear regression for all the clusters as

```
array[H] vector[P+1] betas;
```

and this allowed us to access for `h = 1,...,H` the correspondent `betas[h]` efficiently and to define it using the vectorized version of the normal probability function, i.e.

```
for ( h in 1:H )
    betas[h] ~ normal(mu_0, sigma_0);
```

where with `mu_0` and `sigma_0` we denoted $\boldsymbol{\mu_0}$ and σ_0 respectively, as in (4).

4.2. Simulation Study

To verify the correctness of the implementation and show the ability of the model to correctly estimate parameters and cluster areal locations we performed a study on simulated data. Following [4] we simulated spatio-temporal data from a grid of 10 × 10 areal locations belonging to seven different clusters, as shown in figure 9.

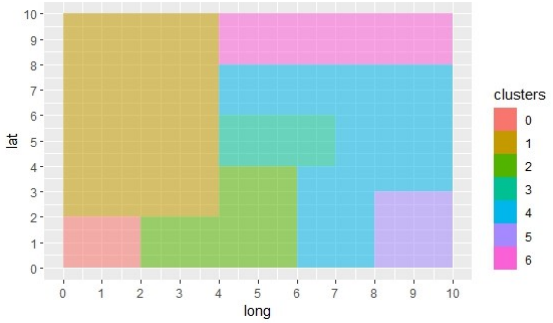
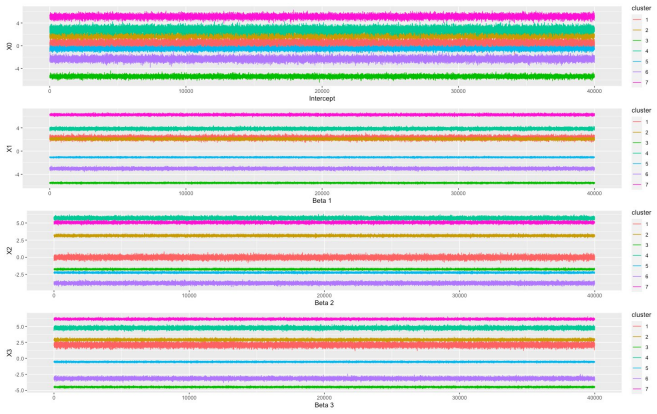
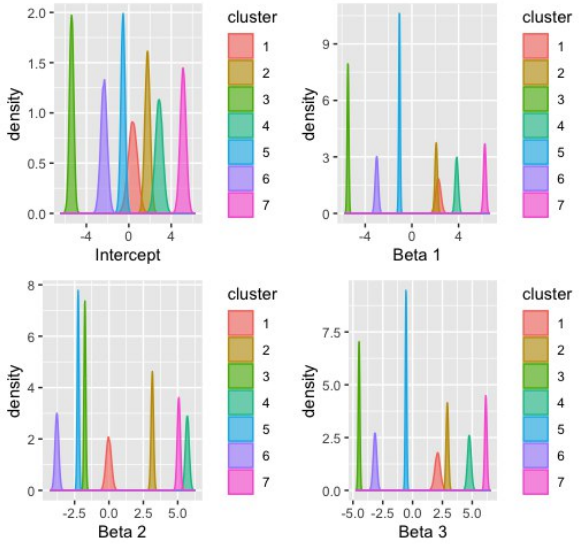


Figure 9: Underlying cluster structure for simulated data.

Each areal location had three covariates and four regression coefficients, the first ones all sampled independently from a normal standard distribution while the second ones indipendently sampled from a multivariate Gaussian distribution with different means according to the cluster to which the areal location belongs. The spatial autocorrelation parameter ρ is set to 0.95, while $\tau^2 = \sigma^2 = 1$. In this case, we used no prior for $\boldsymbol{\mu}_0$ and σ_0^2, set equal to zero and one respectively, as there was no reason to further facilitate the detection of clusters.
We run the MCMC sampler for 20000 iterations and discarded the first 10000 as burn-in and figure 10 shows the posterior inference obtained. The density plots show that the true $\boldsymbol{\beta}$ coefficients are accurately recovered, while the trace plots imply good mixing and convergence.



(a) Trace plots



(b) Estimated density

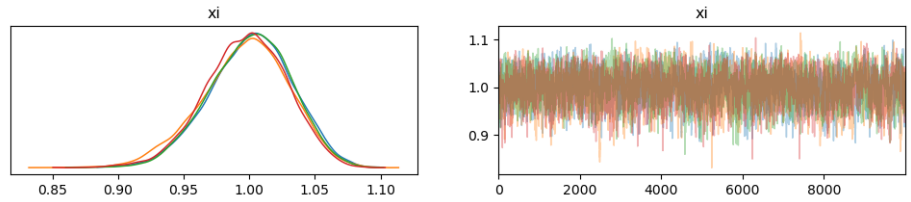
Figure 10: Posterior inference of the cluster-specific regression coefficients $\boldsymbol{\beta}$.

Estimating the clustering with the Binder loss and the VI we obtained the same results and we were able to find all the 7 clusters from which the data came, which still underlines the goodness of the model used.

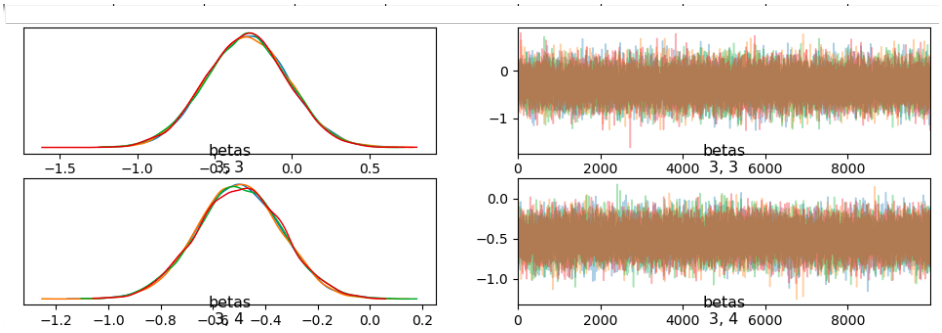
5. Conclusions

In this work we have used a Bayesian model to study the birth rate decrease phenomenon in Italy. In particular, following the literature, we have modeled the spatial correlation between the Italian provinces by means of spatio-temporal random effects with a conditional autoregressive prior. Moreover, we have allowed for clustering of the areal locations through a Dirichlet Process prior on the regression parameters of the predictors.
We have implemented the code of the model in `STAN` putting special emphasis on code optimization as described in section 4.1. After having tested said code on simulated spatio-temporal data, we applied it to the dataset of Italian fertility featured in section 1.1 obtaining insightful results. Indeed, we have recognized six clusters each of which is characterized by different trends of fertility rate and economic variables as employment or inactivity rate. We can conclude that the general trend of Italian fertility rate is, with different performances across different clusters as in section 3.3, strongly negative. Two provinces are an exception to this trend, Gorizia and Bolzano, which present peculiar behaviors.
Furhter work could be developed to enhance the performance of the model, for example we could put a Dirichlet Process prior on the autoregressive coefficient ξ, as in [4], and allow for clustering also by means of these variables. This would, most probably, lead to the implementation of the model outside of `STAN` as a Polya Urn scheme is required without possibility of reformulating the model without it.

A. Appendix A. MCMC Diagnostics



(a) Traceplot of ξ



(b) Traceplot of some components of $\beta_j^\star$

Figure 11: Traceplots for some variables in the state of the MCMC

From figure 11 we can say that the performance of the MCMC are good. Indeed, the values of the samples fluctuate around a stable mean and explore the entire target distribution as they do not get stuck in local modes.

References

[1] Rushworth A, Lee D, and Mitchell R. A spatio-temporal model for estimating the long- term effects of air pollution on respiratory hospital admissions in greater london. *Spatial and Spatio-temporal Epidemiology*, 2014.

[2] Michela Finizio. L'inverno demografico dell'italia: ecco le cause e come superarlo, December 2022. [Online; posted 13-December-2022].

[3] Duncan Lee, Alastair Rushworth, Gary Napier, and William Pettersson. Carbayesst version 3.3: Spatio-temporal areal unit modelling in r with conditional autoregressive priors.

[4] Alexander Mozdzen, Andrea Cremaschi, Annalisa Cadonna, Alessandra Guglielmi, and Gregor Kastner. Bayesian modeling and clustering for spatio-temporal areal data: An application to italian unemployment. *Spatial Statistics*, 2022.

[5] Redazione OpenPolis. Il calo delle nascite dopo l'emergenza covid, May 2022. [Online; posted 31-May-2022].

[6] Stan Development Team. Stan user's guide, 2011–2022. [Online].