Exploratory analysis of accesses to support centers for gender-based violence in Apulia

Data

The dataset employed regards the counts of accesses to gender-based violence support centers in the Apulia region by residence municipality of the women victims of violence in 2021-2024. `R` codes to generate the dataset are in the R script posted here which this report is based on.

Here, we only take into account the violence reports which support centers actually take charge of, at the risk of underestimating the counts of gender-based violence cases. This choice is driven by the need of avoiding duplicated records, since e.g. it may happen that a support center redirects a victim to another support center.

In order to avoid singletons, i.e. municipalities with no neighbours, in the spatial structure of the dataset, the Tremiti Islands need to be removed from the list of municipalities included (0 accesses recorded so far).

Therefore, the municipality-level dataset in scope consists of 256 areas.

We can only take into account the accesses to support centers for which the origin municipality of victims is reported; therefore the total count of accesses in scope is 1477, 1516, 1822 and 1778for the reference years respectively:

Here, we plot the log-access rate per residence municipality, i.e. the logarithm of the ratio between access counts and female population. Blank areas correspond to municipalities from which zero women accessed support centers (82 municipalities).

Covariates

Our target is explaining the number of accesses to support centers, y, defined at the municipality level, on the basis of a set of candidate known variables. y is modelled with simple Poisson regression.

We have at disposal a number of candidate explanatory variables, which include the distance of a municipality from the closest support center and a set of variables measuring social vulnerability under different dimensions; these latter covariates are provided by the ISTAT and are among the components of the Municipality Frailty Index (MFI). A more detailed description of MFI components is in this excel metadata file.

All covariates are quantitative variables and to ease model interpretation are scaled to have null mean and unit variance.

- TEP_th, i.e. the distance of each municipality from the closest municipality hosting a support center. Distance is measured by road travel time in minutes (acronym TEP stays for Tempo Effettivo di Percorrenza, i.e. Actual Travel Time). Since to the best of our knowledge the list of active support centers changed between 2022 and 2023, we employ the list of centers active until 2022 for 2021-2022 data, and the list of centers active in 2023 for 2023 data.
- AES, the distance from the closest infrastructural pole, always measured in travel time. Infrastructural poles are defined as municipalities or clusters of neighbouring municipalities provided with a certain endowment in health, education and transport infrastructure (more details can be found, e.g., here).
- MFI, i.e. the decile of municipality frailty index.

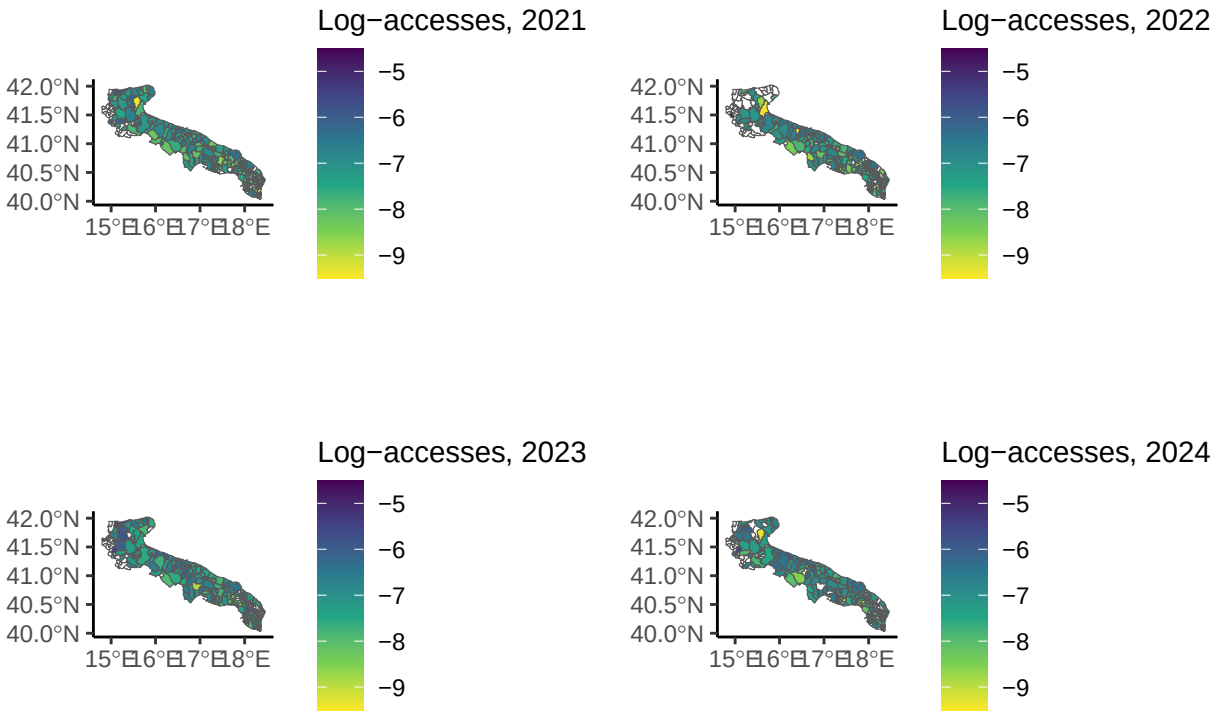

Figure 1: Log-access rate

- PDI, i.e. the dependency index, i.e. population either ≤ 20 or ≥ 65 years over population in $[20-64]$ years.
- ELL, i.e. the proportion of people aged $[25-54]$ with low education.
- ERR, i.e. employment rate among people aged $[20-64]$.
- PGR, i.e. population growth rate with respect to 2011.
- UIS, i.e. the ventile of the density of local units of industry and services (where density is defined as the ratio between the counts of industrial units and population).
- ELI, i.e. the ventile of employees in low productivity local units by sector for industry and services.

First, we visualise the correlations among these explanatory variables:

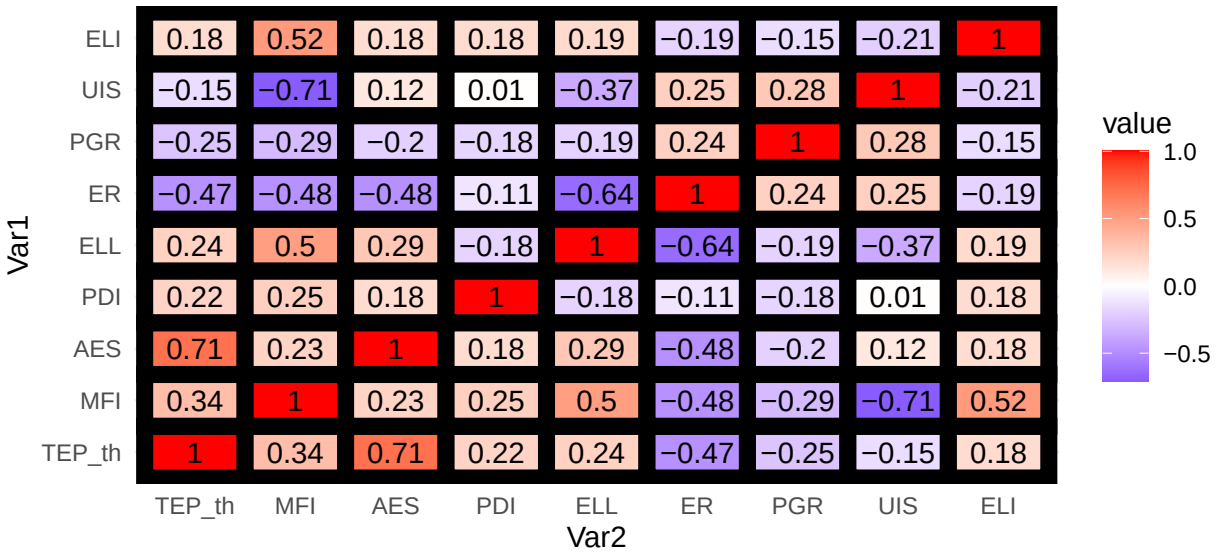

Figure 2: Correlations in explanatory variables

We see the correlation between the two distances is very high (0.72), and so is the correlation between the fragility index decile and the density of productive units.

In the first case, we drop the distance from the nearest infrastructural pole. In the latter we drop `MFI`, which is a combination of all covariates except for `TEP_th`, and is a weakly informative choice.

Nonspatial regression

We regress the counts of accesses y to support centers on the aforementioned explanatory variables. To interpret regression coefficients in an easier way, all covariates are scaled to zero mean an unit variance.

$$y_{it} \mid \eta_{it} \sim \text{Poisson}(E_{it}\, e^{\eta_{it}}) \quad \text{where} \quad \eta_{it} = X_{it}^{\top}\beta \tag{1}$$

Where X are the covariates defined earlier, β are covariate effects, and E_{it} is the female population aged ≥ 15 in municipality i and year t.

To gain more insight on the role of all explanatory variables we show the posterior summaries of the regression model; the posterior distribution is approximated with the INLA (Van Niekerk et al. 2023).

% latex table generated in R 4.5.0 by xtable 1.8-4 package % Tue Aug 12 13:26:11 2025

Effect	Mean	Sd	0.025quant	0.975quant
Int_2021	-7.333	0.030	-7.392	-7.274
Int_2022	-7.305	0.030	-7.364	-7.247
Int_2023	-7.128	0.028	-7.183	-7.073
Int_2024	-7.152	0.028	-7.208	-7.097
TEP_th	-0.227	0.017	-0.260	-0.195
ELI	-0.088	0.015	-0.117	-0.059
PGR	0.017	0.020	-0.021	0.056
UIS	0.040	0.016	0.009	0.071
ELL	-0.110	0.021	-0.150	-0.070
PDI	-0.065	0.021	-0.106	-0.025
ER	-0.212	0.022	-0.256	-0.169

- `TEP_th_22`: The distance from the closest support center appears to play an important role. The easiest interpretation is that the physical distance represents a barrier to violence reporting. This is quite intuitive if we think of the material dynamics of reporting gender-based violence: one could reasonably expect violent men to prevent their partners to take a long rout and come out to report the violence suffered.
- `ELI`: The (ventile of the distribution of the) share of employees in low productivity economic units is a clear indicator of (relative) economic underdevelopment. The most naive interpretation would be that in underdeveloped areas reporting gender violence is somewhat harder than in developed ones.
- `PGR`: The association with population growth rate does not appear to be significantly different from zero.
- `UIS`: The (ventile of the distribution of the) density of production units does not appear to have a significantly $\neq 0$ association with VAW reporting.
- `ELL`: The association with the proportion of people with low educational level has negative sign and is high in absolute value. The interpretation seems quite easy: cultural development, in general, would encourage reporting violence.
- `PDI`: The association with population dependency index is negative, and its interpretation is ambiguous as well since the proportion of old or very young people can be read as a limiting factor both for VAW incidence and reporting.
- `ER`: The association with employment rate is very strong and bears negative sign for 2021 and 2023 data.

Spatial regression

Exploratory analysis of residuals We plot the log-residuals ε of the nonspatial regression models, defined as $\varepsilon := \ln y_{it} - \ln \hat{y}_{it}$ being $\hat{y}_{it}$ the fitted value.

Residuals may exhibit spatial structure. To assess it, we employ the Moran and Geary tests. Since

Please notice that log-residuals only take finite values across the municipalities whose female citizens have reported at least one case of violence.

Additionally, this set of municipalities may include some singletons, which we remove to assess the value of the Moran and Geary statistics. Thus, for each year we have defined the indexes set `nonzero_con` as the set of municipalities from which at least one case of gender-based violence has been reported, *and* which have at least one neighbouring municipality from which at least one case of gender-based violence was reported as well. For brevity, we only show the standardised I values, which under the null hypothesis should be distributed as $N(0, 1)$. The Geary's test is also included for completeness.

% latex table generated in R 4.5.0 by xtable 1.8-4 package % Tue Aug 12 13:26:13 2025

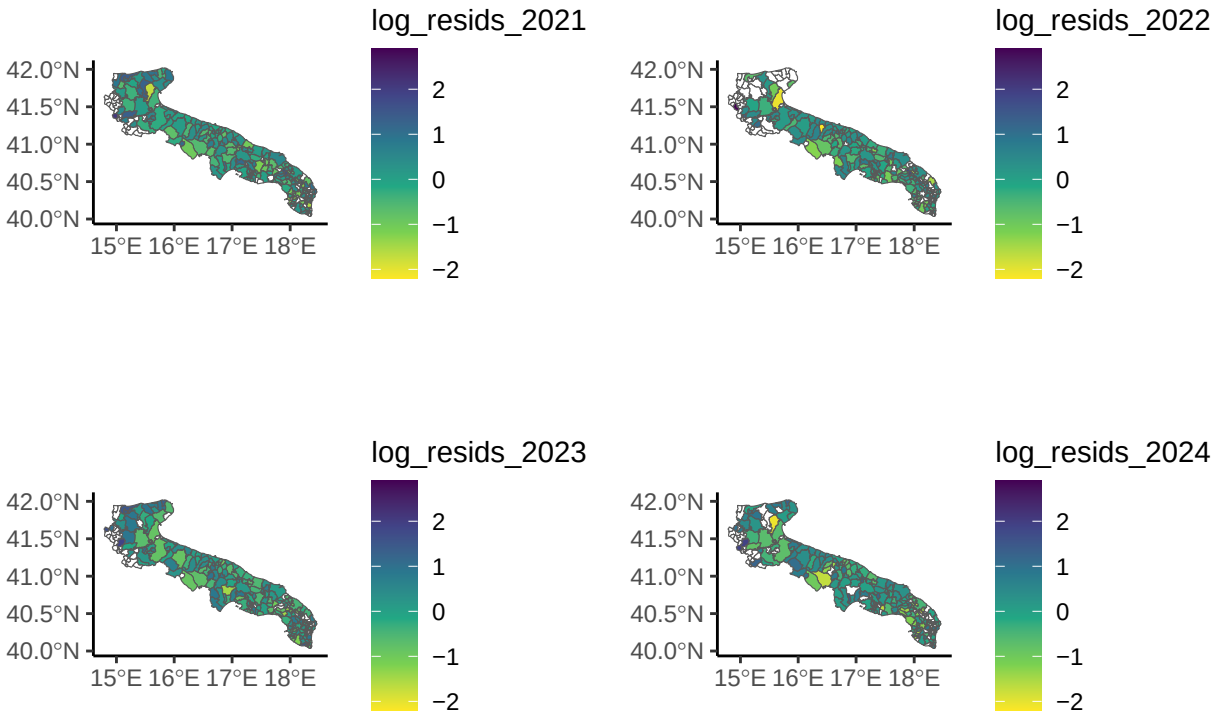

Figure 3: Log-residuals in GLM regression

Year	Test	Statistic_std	p.value
2021	Moran	1.78	0.04
2021	Geary	2.18	0.01
2022	Moran	2.20	0.01
2022	Geary	2.37	0.01
2023	Moran	4.85	0.00
2023	Geary	4.93	0.00
2024	Moran	3.94	0.00
2024	Geary	4.30	0.00

We find evidence for spatial autocorrelation. However, we must stress out that this result does not refer to all the regional territory, but only to a subset of all municipalities.

Based on the autocorrelation evidence, though it has only been assessed for a subset of all municipalities, we try implementing some simple spatial models by adding a conditionally autoregressive latent effect, say z, to the linear predictor

$$\eta_{it} = X_{it}^{\top}\beta + z_{it} \tag{2}$$

We test a total of four models, all of which have a prior distribution depending on the spatial structure of the underlying graph, in this case the Apulia region.

In the following, the area-specific latent field is denoted as $z_i = (z_{i,2021}\, z_{i,2022}\, z_{i,2023}, z_{i,2024})^{\top}$

We describe the spatial structure starting from municipalities neighbourhood, and introduce the neighbourhood matrix W, whose generic element w_{ij} takes value 1 if municipalities i and j are neighbours and 0 otherwise. For each $i \in [1, n]$, $d_i := \sum_{j=1}^{n} w_{ij}$ is the number of neigbours of i-th municipality. Please notice we have have $n = 256$.

For all models, we define Σ as the covariance matrix of the latent effect z; its off-diagonal entries describe dependence in z between different years. Spatial models are computed by approximating the marginal posteriors of interest via the Integrated Nested Laplace Approximation (INLA), adopting the novel Variational Bayes Approach (Van Niekerk et al. 2023). Before proceeding with model structure, we need a short digression on how `R-INLA` works. The INLA framework is hierarchical and is based on a distinction between latent Gaussian fields (GFs) and hyperparameters. In our case, the former include intercepts, covariate effects and the spatial latent field. In the INLA literature latent Gaussian fields are frequently labelled as x but we avoid doing so in order not to introduce notation ambiguity with respect to covariates. Hyperparameters in our case coincide with the parameters of z. Then, it is necessary to approximate the posterior distribution of both hyperparameters and latent GFs.
INLA follows a framework of numerical integration over a grid of hyperparameters, which is centered at the mode. Locating the posterior mode of hyperparameters is thus the first, crucial task to approximate all posterior distributions involved in the model. `R-INLA` employs an exploration strategy over the whole real axis (Rue, Martino, and Chopin 2009). For this reason, hyperparameters need be re-parametrised in such a way to be defined over all $\mathbb{R}$, and the priors are thus actually written on the so-called internal scale of hyperparameters. This approach may be hiding a trick if specifical constraints are required on hyperparameters. For instance, Σ needs to be positive-definite and symmetric. If the hyperparameter exploration is carried out on the whole real axis, nothing ensures, in principle, that the values of the single entries in Σ ensure it is positive definite. For instance, consider the case the hyperparameter exploration procedure comes across this correlation matrix

$$\Sigma_{\text{ND}} \begin{pmatrix} 1 & 0.4 & 0.9 \\ 0.4 & 1 & -0.4 \\ 0.9 & -0.4 & 1 \end{pmatrix}$$

It can be seen that Σ_{ND} has one negative eigenvalue, and is thus indefinite: it cannot be a valid correlation matrix and would cause trouble if, e.g., the covariance matrix is assigned a Wishart prior. A trivial solution

is to define the correlation matrix as a quadratic form, say $\Sigma := BB^\top$, and setting a prior distribution on B. Of course, it is convenient to define triangular B matrix - otherwise, both a high number of parameters would be required, and the prior on Σ could not be calculated (could it?) starting from the prior on B as the Jacobian matrix would *not* be squared.

Priors for spatial effects and hyperparameters are defined in this R script. The general structure follows the framework of the `bigDM` R package (Vicente et al. 2023).

ICAR model The Intrinsic CAR model is the simplest formulation among spatial autoregressive models. The conditional distribution of each value $z_i \mid z_{-i}$ is:

$$z_i \mid z_{-i} \sim \mathcal{N}\left(\sum_{j=1}^{n} \frac{w_{ij}}{d_i} z_j \,, \frac{1}{d_i}\Lambda^{-1}\right) \tag{3}$$

And the joint prior distribution is:

$$z \mid \Sigma \sim \mathcal{N}\left(0, \Sigma \otimes (D - W)^{+}\right) \tag{4}$$

Where z is a 256×4 matrix. Since the joint distribution of z is improper, a sum-to-zero constraint, i.e. $\sum_{i=1}^{n} z_{it} = 0$ for $t = 2021, 2022, 2023, 2024$ is required for identifiability.

PCAR model The intrinsic autoregressive model is relatively simple to interpret and to implement, while also requiring the minimum number of additional parameters (either the scale or the precision).

The drawback, however, is that we implicitly assume a deterministic spatial autocorrelation coefficient equal to 1. When the autocorrelation is weak, setting an ICAR prior may be a form of misspecification.

A generalisation of this model is the PCAR (proper CAR), which introduces an autocorrelation parameter ρ. We assign a penalised complexity (PC) prior to ρ, and, following the thumb-rule proposed by (Simpson et al. 2017) and (Riebler et al. 2016) for the Besag-York-Mollié model (see after) we assume *a priori* that $\pi(\rho \leq \frac{1}{2}) = \frac{2}{3}$.

$$z_i \mid z_{-i} \sim \mathcal{N}\left(\sum_{j=1}^{n} \rho \frac{w_{ij}}{d_i} z_j \,, \frac{1}{d_i}\Lambda^{-1}\right) \tag{5}$$

We show the posterior summary for the autocorrelation coefficient, obtained with the data at hand.

```
##       mean         sd quant0.025  quant0.25   quant0.5  quant0.75 quant0.975
##  0.8496518  0.0698914  0.6808226  0.8107266  0.8615526  0.9011834  0.9500745
```

The credible interval for ρ is quite pushed towards unity, denoting the model estimates a strong spatial autocorrelation.

Leroux model As an alternative to take into account both structured and unstructured latent effects, we also test the Leroux autoregressive model (Leroux, Lei, and Breslow 2000); throughout this report, this model will be referred to as the LCAR. In this case, the local prior for z_i is

$$z_i \mid z_{-i} \sim \mathcal{N}\left(\sum_{j=1}^{n} \frac{\xi w_{ij}}{1 - \xi + \xi d_i} z_j \,, \Lambda^{-1} \frac{1}{1 - \xi + \xi d_i}\right) \tag{6}$$

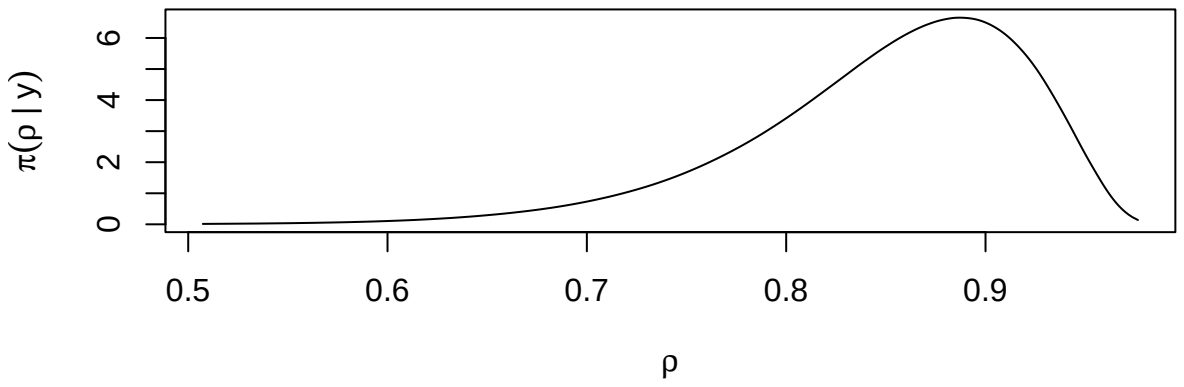

Figure 4: Posterior marginal of the PCAR autocorrelation parameter

Where $\xi \in [0, 1]$ is labelled as the mixing parameter. A more interesting representation of the Leroux model is the joint prior

$$z \mid \Lambda, \xi \sim \mathcal{N}(0, [\Lambda \otimes (\xi L + (1 - \xi)I)]^{-1})$$

where $L := D - W$ is the graph Laplacian matrix, W is the neighbourhood matrix and D is the corresponding degree matrix. We can clearly see how the mixing parameter allocates variability between two precision components, i.e. the Laplacian matrix for the spatial part and the the identity matrix for the noise.

The drawback of this model is the scarce interpretability with respect to more sophisticated ones like the BYM.

BYM model Another popular model to control for both spatial autocorrelation and random noise is the Besag, York and Mollié model. We employ a simplified formulation, with a unique mixing parameter for all three years. Under this model the latent effect is defined as:

$$z = \sqrt{\phi} u M + \sqrt{1 - \phi} v M \tag{7}$$

Where:

- $u \sim \mathcal{N}(0, I_p \otimes L_{\text{scaled}})$ is an independent multivariate ICAR process whose precision matrix is scaled in order that the geometric mean of marginal variances (i.e. the diagonal entries of L^+) is one
- $v \sim \mathcal{N}(0, I_p \otimes I_n)$ is a Standard Normal variable
- M is a positive definite matrix such that $M'M = \Lambda^{-1}$. It is not necessarily the Cholesky factor of the scale parameter; in fact, a convenient but not unique way to define it may be $M = D^{-\frac{1}{2}} E^\top$ where E and D are the eigenvector and eigenvalues matrices of Λ (Urdangarin et al. 2024).
- $\phi \in [0, 1]$ is the mixing parameter.

Please notice that in the case of our data, $p = 3$ and $n = 256$. We assign a Uniform prior on ϕ (but the PC-prior would be a more rigorous choice) and the usual Wishart prior to Λ.

The BYM model features a variance mixing parameter, whose interpretation is eased by allowing to scale the ICAR and IID components.

We briefly compare the three models in scope through the WAIC (Gelman, Hwang, and Vehtari 2014):

% latex table generated in R 4.5.0 by xtable 1.8-4 package % Tue Aug 12 13:28:41 2025

Model	WAIC	Eff_params
Null	4798.741	44.505
ICAR	3899.529	257.647
PCAR	3918.801	265.977
Leroux	3907.826	265.791
BYM	3880.151	257.395

As we can see, adding a spatial model is an improving element, not a waste of complexity.

Here we show some posterior summaries for β under the BYM model.

% latex table generated in R 4.5.0 by xtable 1.8-4 package % Tue Aug 12 13:28:41 2025

Mean	Sd	Q_0.025	Q_0.975
-7.439	0.049	-7.537	-7.344
-7.613	0.061	-7.737	-7.496
-7.199	0.048	-7.295	-7.105
-7.339	0.054	-7.447	-7.234
-0.204	0.046	-0.295	-0.113
-0.001	0.039	-0.078	0.077
0.077	0.043	-0.008	0.163
0.086	0.043	0.002	0.171
-0.228	0.052	-0.330	-0.126
-0.036	0.047	-0.128	0.056
-0.274	0.061	-0.393	-0.156

Estimations of β differ slightly from the nonspatial model and credibility intervals are generally wider due to increased uncertainty.

- `TEP_th_22` The effect of the distance from the closest support center remains similar in mean and the interpretation is not altered.
- `ELI`: The effect of the incidence of low-productivity economic units is utterly negligible
- `PGR`: The association with population growth rate can only be considered barely significant for 2022 data
- `UIS`: The association with the density of productive units is negligible for 2021 and 2022 data, and can be considered slightly significant for 2023, bearing positive sign.
- `ELL`: The association with the incidence of low education levels, is even higher in mean than under the nonspatial model. We interpret this result as a strong *potential* impact of education on the chance that gender violence is reported
- `PDI`: The effect of structural dependency index is utterly negligible, as for the GLM.
- `ER`: The effect associated with employment rate is increased for 2021 data, more than doubled for 2022 data, and slightly increased for 2023 data. How to interpret this finding? Employment rate is clearly an indicator of economic development, hence the easiest interpretation is that - as it was with `ELI` under the nonspatial model - in more developed areas there is a higher chance that gender violence is reported.

We show the expected latent effects:

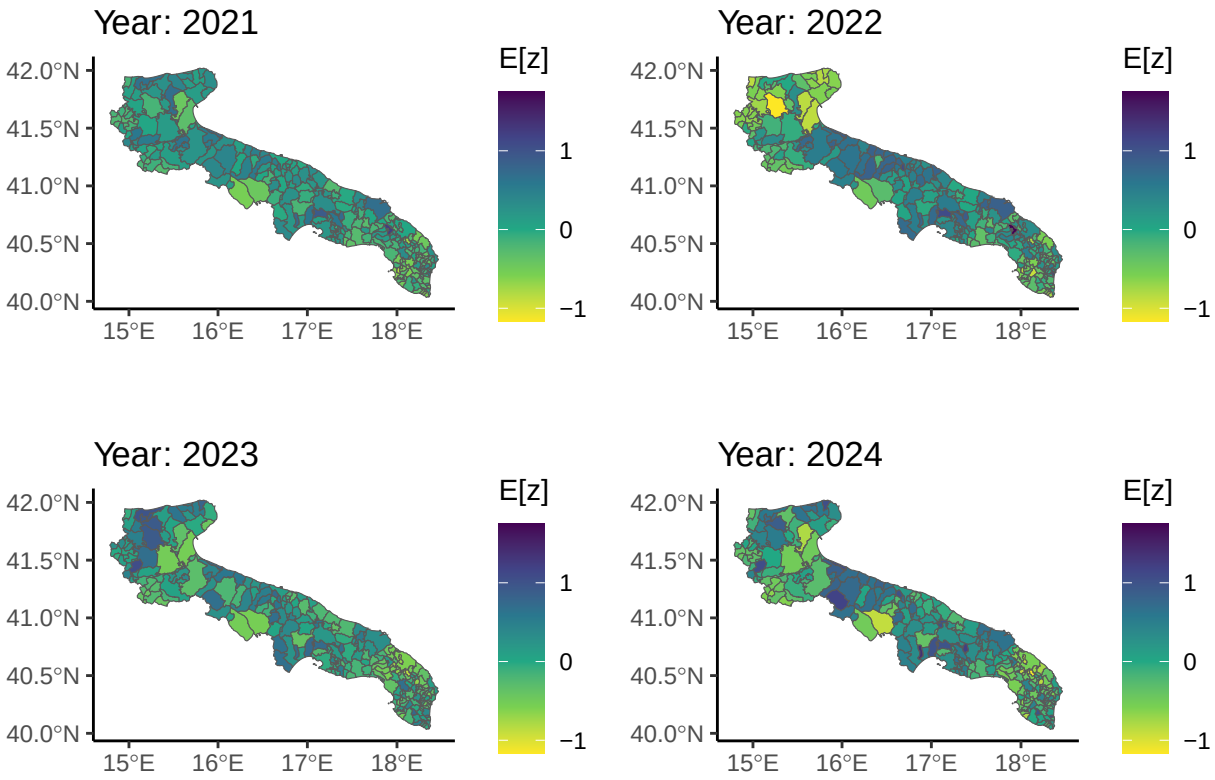

Figure 5: Estimated BYM latent effect

Gelman, Andrew, Jessica Hwang, and Aki Vehtari. 2014. "Understanding Predictive Information Criteria for Bayesian Models." *Statistics and Computing* 24 (6): 997–1016. https://doi.org/10.1007/S11222-013-9416-2.

Leroux, Brian G., Xingye Lei, and Norman Breslow. 2000. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, edited by M. Elizabeth Halloran and Donald Berry, 179–91. New York, NY: Springer New York. https://doi.org/https://doi.org/10.1007/978-1-4612-1284-3_4.

Riebler, Andrea, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. 2016. "An Intuitive Bayesian Spatial Model for Disease Mapping That Accounts for Scaling." *Statistical Methods in Medical Research* 25 (4): 1145–65.

Rue, Haavard, Sara Martino, and Nicholas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society Series B: (Methodological)* 71 (2): 319–92. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Simpson, Daniel, Haavard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28. https://doi.org/10.1214/16-STS576.

Urdangarin, Arantxa, Tomas Goicoa, T. Kneib, and M. D. Ugarte. 2024. "A Simplified Spatial+ Approach to Mitigate Spatial Confounding in Multivariate Spatial Areal Models." *Spatial Statistics* 59: 100804. https://doi.org/10.1016/j.spasta.2023.100804.

Van Niekerk, Janet, Elias Krainski, Denis Rustand, and Haavard Rue. 2023. "A New Avenue for Bayesian Inference with INLA." *Computational Statistics and Data Analysis* 181. https://doi.org/10.1016/j.csda.2023.107692.

Vicente, Gonzalo, Aritz Adin, Tomás Goicoa, and María Dolores Ugarte. 2023. "High-Dimensional Order-Free Multivariate Spatial Disease Mapping." *Statistics and Computing* 33 (5): 104.