Explorative 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 during 2022. `R` codes to generate the dataset are in the R script posted here which this report is based on.

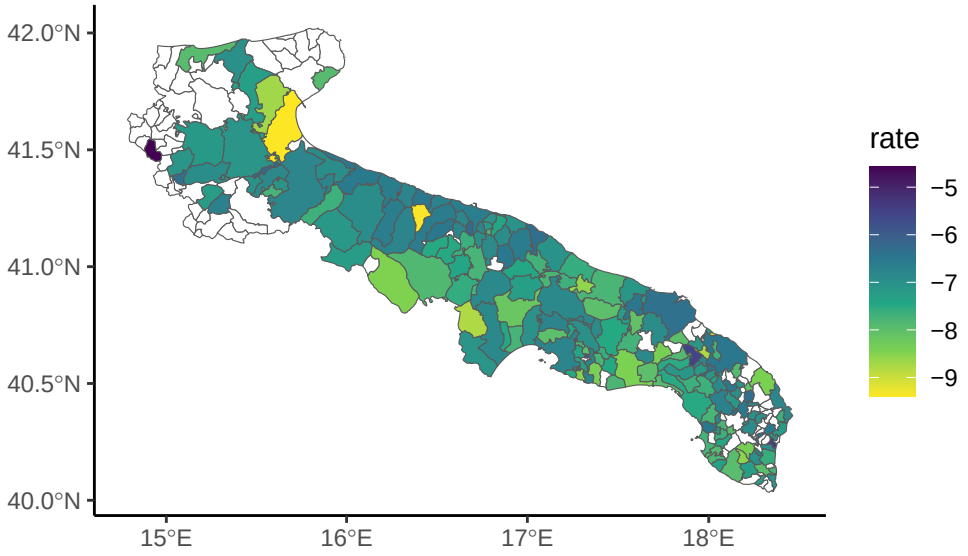
Here, we only take into account the violence reports which support centers actually take charge of, at the risk of underestimating the counts of gender-based violence cases. This choice is driven by the need of avoiding duplicated records, since e.g. it may happen that a support center redirects a victim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

In order to avoid singletons in the spatial structure of the dataset, we removed the Tremiti Islands from the list of municipalities included (0 accesses to support centers in 2022).

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

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 2259. Among these accesses, 1516 were taken charge of.

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.

We model y via a simple Poisson GLM.

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.A more detailed description of these covariates is in this excel metadata file.

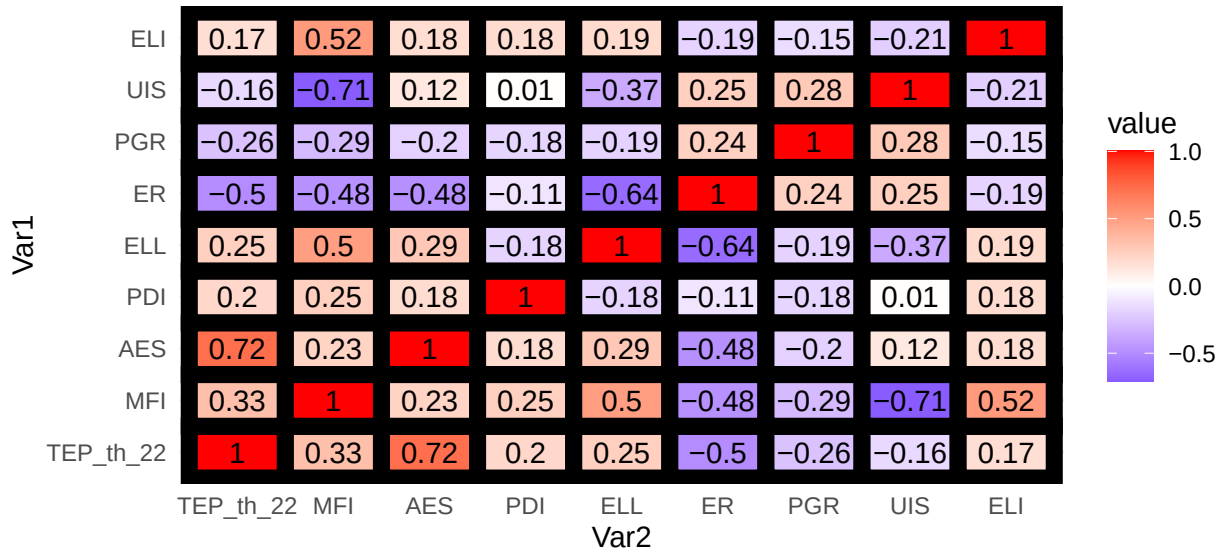
All covariates are scaled to have null mean and unit variance.

- TEP, 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).

For instance, the support center designated for the municipality of Adelfia (province of Bari, 3rd municipality in the dataset) is located in Capurso (BA). Then, TEP_3 denotes the travel time between Adelfia and Capurso (17 minutes).

- AES, the distance from the closest infrastructural pole, always measured in travel time.
- MFI, i.e. the decile of municipality vulnerability index.
- 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:



Then, we implement a very simple forward selection algorithm. At each iteration, we add to the model the covariate allowing for the lowest BIC, until adding an additional covariate does not allow to reduce it anymore:

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 We do so because, if taken alone, the distance from the closest support center appears a slightly better predictor, using the Schwarz information criterion (or, indifferently, the Akaike Information Criterion):

```
stats::BIC(glm(N_ACC ~ 1 + AES, family = "poisson",
               offset = log(nn), data = dd_con))
```

```
## [1] 1124.922
```

```
stats::BIC(glm(N_ACC ~ 1 + TEP_th_22, family = "poisson",
               offset = log(nn), data = dd_con))
```

```
## [1] 1120.389
```

We should do the same for the other couple of variables but since `MFI` is a combination of all covariate except for `TEP_th`, we will drop the synthetic indicator and leave the remainder.

```
covariates <- colnames(X)[-c(1, which(colnames(X) %in% c("AES", "MFI")))]
# Covariates included:
covs.in <- c()
# Covariates not included:
BIC.min <- c()
while(length(covs.in) < length(covariates)){
  covs.out <- covariates[which(!covariates %in% covs.in)]

  BICs <- c()
  # At each iteration, we add one of the remaining covariates
  for(j in c(1:length(covs.out))) {
    formula.temp <- paste0("N_ACC ~ 1 + offset(log(nn)) +",
                           paste(covs.in, collapse = "+"),
                           "+", covs.out[j])
    mod.tmp <- glm(formula.temp, data = dd_con, family = "poisson")
    BICs[j] <- stats::BIC(mod.tmp)
  }
  # Covariate allowing for the best model is added:
  BIC.min <- c(BIC.min, min(BICs))
  covs.in <- c(covs.in, covs.out[which.min(BICs)])
}
```

The optimal number of covariates (covariates in the model with minimum BIC) is four:

```
covs.in[c(1:which.min(BIC.min))]
```

```
## [1] "TEP_th_22" "ELI"       "PGR"       "UIS"
```

However, a model with up to 4 covariates would have all significant regression coefficients:

```
summary(glm(
  as.formula(paste0("N_ACC ~ 1 +", paste(covs.in[c(1:5)], collapse = " + "))),
  family = "poisson",offset = log(nn), data = dd_con))$coefficients
```

```
##                Estimate Std. Error     z value     Pr(>|z|)
## (Intercept) -7.44858456 0.04273029 -174.316269 0.000000e+00
## TEP_th_22   -0.40674770 0.04227726   -9.620957 6.522178e-22
## ELI         -0.06083398 0.03318179   -1.833354 6.674988e-02
## PGR          0.12904830 0.04156451    3.104771 1.904261e-03
## UIS         -0.11049111 0.03321299   -3.326744 8.786702e-04
## ELL         -0.06642470 0.03378243   -1.966250 4.926976e-02
```

- The distance from the closest support center seems to play the key 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 come out and report the violence suffered.
- The asociation with population growth rate is harder to interpret. This association is most likely influenced by several demographic instrumental variables we are not keeping into account and would indeed deserve a more dedicated focus.
- The density of production units can be hither considered as a proxy for economic development. The most naive interpretation wuld be that in underdeveloped areas reporting gender violence is somewhat harder than in developed ones.
- The association with the proportion of people with low educational level appears lower. However, the interpretation seems quite easy: cultural development, in general, would encourage reporting violence. Still, let us remind that β_{ELL} is in $\times(10^{-2})$.
- The interpretation of β_{ELI} should, intuitively, be similar to β_{UIS}. We have kept both variables due to their correlation not being particularly high.

With more covariates, no additional valuable association can be found.

In the remainder of this work, we will focus on the two covariates $d, \text{PGR}, \text{ELI}, \text{UIS}$

Nonspatial regression

We regress the counts of accesses y to support centers on the distance from the former. Both covariates are scaled to zero mean an unit variance.

$$y_i \mid \eta_i \sim \text{Poisson}(e^{\eta_i - P_i}) \quad \text{where} \quad \eta_i = \beta_0 + \beta_{\text{TEP}}\text{TEP}_i + \beta_{PGR}\text{PGR}_i + \beta_{ELI}\text{ELI}_i + \beta_{UIS}\text{UIS}_i \tag{1}$$

Where P_i is the female population aged ≥ 15 in municipality i.

```
##
## Call:
## glm(formula = N_ACC ~ 1 + TEP_th_22 + ELI + PGR + UIS, family = "poisson",
##     data = dd_con, offset = log(nn))
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -7.45252    0.04274 -174.355  < 2e-16 ***
## TEP_th_22   -0.39986    0.04173   -9.582  < 2e-16 ***
## ELI         -0.08650    0.03026   -2.859 0.004256 **
## PGR          0.15204    0.03962    3.838 0.000124 ***
## UIS         -0.09545    0.03222   -2.963 0.003051 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 729.12  on 255  degrees of freedom
## Residual deviance: 508.14  on 251  degrees of freedom
## AIC: 1084.7
##
## Number of Fisher Scoring iterations: 5
```

How do we interpret the regression coefficients? Keeping in mind we are working on the logarithm of the access rate, the standard deviation of the distance, expressed in minutes, is:

```
# Distance from closest support center
attr(scale(dists_th_22$TEP_th_22), "scaled:scale")
```

```
## [1] 14.10021
```

Hence e.g. each $14'6''$ of distance of the a given municipality from the closest support center are associated with a decrease of 0.399 units in the log-frequency at which women from that municipality access to support centers.

Spatial regression

We plot the log-residuals ε of the regression model in equation 1, defined as $\varepsilon := \ln y_i - \ln P_i - \ln \hat{y}_i$ being $\hat{y}_i$ the fitted value.

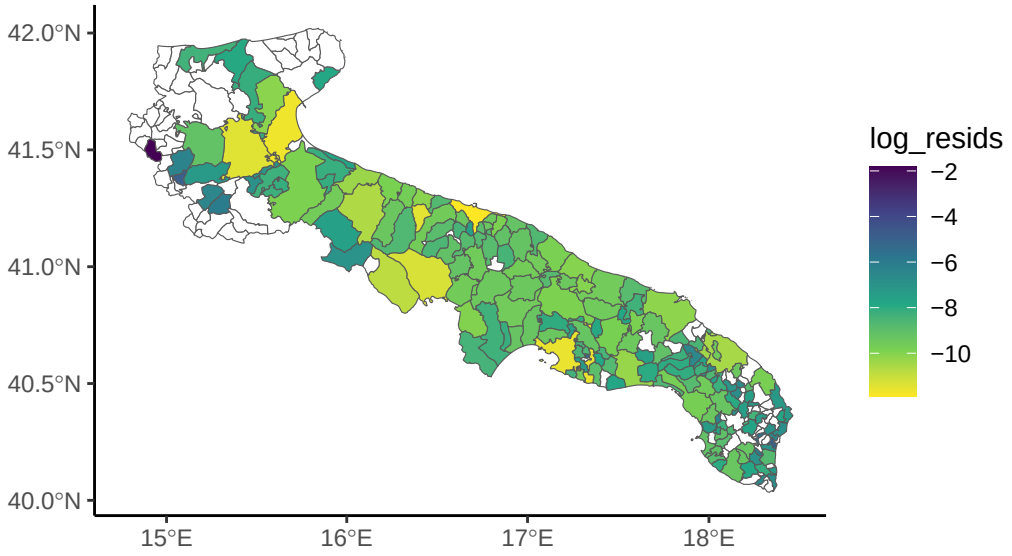


Figure 1: Log-residuals of glm regression using theorical distance as explanatory variable

```
## Warning in spdep::poly2nb(dd_con[nonzero_con, ]): neighbour object has 2 sub-graphs;
## if this sub-graph count seems unexpected, try increasing the snap argument.
```

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 175 municipalities whose female citizens have reported at least one case of violence in 2022.

Additionally, this set of municipalities includes 2 singletons, which we remove to assess the value of the Moran and Geary statistics. Thus, 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 municipalities from which at least one case of gender-based violence was reported.

```
spdep::moran.test(resids_glm_th[nonzero_con],
                  listw = spdep::nb2listw(nb_con_nonzero))
```

```
##
##  Moran I test under randomisation
##
## data:  resids_glm_th[nonzero_con]
## weights: spdep::nb2listw(nb_con_nonzero)
##
## Moran I statistic standard deviate = 6.1962, p-value = 2.892e-10
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.329797541      -0.005813953       0.002933765
```

```
spdep::geary.test(resids_glm_th[nonzero_con],
                  listw = spdep::nb2listw(nb_con_nonzero))
```

```
##
##  Geary C test under randomisation
##
## data:  resids_glm_th[nonzero_con]
## weights: spdep::nb2listw(nb_con_nonzero)
##
## Geary C statistic standard deviate = 4.551, p-value = 2.669e-06
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##        0.71991557        1.00000000       0.00378755
```

In both cases, we find evidence for spatial autocorrelation. However, we must stress out this result does not refer to all the regional territory, but only to a subset of all municipalities (173 over 257)

Based on the autocorrelation evidence, though it has only been assessed for a subset of all municipalities, we try implementing some simple conditional autoregressive models by augmenting the linear predictor

$$\eta_i = \beta_0 + \beta_d d_i + \beta_{PGR}\text{PGR}_i + \beta_{ELI}\text{ELI}_i + \beta_{UIS}\text{UIS}_i + z_i \tag{2}$$

Where z_i represents the latent Gaussian effect. We test three models, all of which have a prior distribution depending on the spatial structure of the underplying graph, in this case the Apulia region.

We describe the spatial structure starting from municipality 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. Plase notice we have have $n = 256$.

For all models, we define σ^2 as the scale parameter of the latent effect, and in order to avoid overfitting we set a PC-prior on it with rate parameter $\lambda = 1.5$, such that $\text{Prob}(\sigma > \lambda) = 0.01$

Spatial models are computed by approximating the marginal posteriors of interest via the Integrated Nested Laplace Approximation (INLA), adopting the novel Variational Bayes Approach **?**.

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

$$z_i \mid z_{-i} \sim N\left(\sum_{j=1}^{n} \frac{w_{ij}}{d_i} z_j \,,\, \frac{\sigma^2}{d_i}\right) \tag{3}$$

The remainder of the model follows the same notation as eq. 1.

PCAR model The intrinsic autoregressive model is relatively simple to interpret and to implement, while also requiring the minimum number of additional parameter (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 α:

$$z_i \mid z_{-i} \sim N\left(\sum_{j=1}^{n} \alpha \frac{w_{ij}}{d_i} z_j \,,\, \frac{\sigma^2}{d_i}\right) \tag{4}$$

The `R` code to implement the PCAR model in `R-INLA` is in Appendix.

We show the posterior summary for the autocorrelation coefficient. It does not appear strong indeed. The credible interval also appears uncannily wide:

```
## Mean            0.655696
## Stdev           0.190775
## Quantile  0.025 0.232785
## Quantile  0.25  0.527872
## Quantile  0.5   0.686774
## Quantile  0.75  0.808034
## Quantile  0.975 0.933718
```

BYM model Perhaps, our data are generated by a process dominated from noise. We can thus try a different path: the BYM model. On a preliminary stance, we keep trusting in the accuracy of the Laplace approximation and stick to INLA. On a later stage, it would be more rigorous to compare INLA results to the posteriors of a model estimated with MCMC.

The BYM model we employ follows the parametrisation of (Riebler et al. 2016):

$$z_i = \sigma\left(\sqrt{\phi}u_i + \sqrt{1-\phi}v_i\right) \tag{5}$$

where u is an ICAR field, v is an IID standard Gaussian white noise i.e. $v \sim N(0, I)$, and ϕ is a mixing parameter $\in [0, 1]$.

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). In this case, the local prior for z_i is

$$z_i \mid z_{-i} \sim N\left(\sum_{j=1}^{n} \frac{\xi}{1-\xi+\xi d_i}\frac{w_{ij}}{d_i} z_j\,,\, \frac{\sigma^2}{1-\xi+\xi d_i}\right) \tag{6}$$

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

$$z \mid \sigma^2, \xi \sim N(0, \sigma^2[\xi R + (1-\xi)I]^{-1})$$

where $R := D - W$ is the graph Laplacian matrix, W is the neighbourhood matrix and D is the corresponding degree matrix.

Comparison We briefly compare these models through the WAIC (Gelman, Hwang, and Vehtari 2014):

```
## # A tibble: 4 x 3
##   Model   WAIC Eff_Params
##   <chr>  <dbl>      <dbl>
## 1 ICAR    962.       76.4
## 2 PCAR    955.       78.1
## 3 BYM     950.       76.2
## 4 Leroux  952.       76.3
```

As we can see, models taking into account random noise have a better performance. Please notice the effective number of parameters, i.e. the number of *unconstrained* parameters is higher in the ICAR than in the BYM and Leroux models, even though these require an additional parameter.

We show the postetior summary of covariates effects under the BYM model. There are no noteworthy differences with other models in terms of covariate effects estimation.

```
##                mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -7.603 0.065     -7.733   -7.602     -7.479 -7.602   0
## TEP_th_22   -0.370 0.076     -0.519   -0.371     -0.221 -0.371   0
## UIS         -0.089 0.063     -0.211   -0.090      0.036 -0.090   0
## PGR          0.138 0.070      0.002    0.138      0.275  0.138   0
## ELI         -0.063 0.063     -0.185   -0.063      0.061 -0.063   0
```

The interpretation the posteriors of β under this spatial model is not exactly the same as for the GLM. As we see, while the distance from the closest support center remains a strong predictor, the two economic variables seem to lose their significance; this is mainly due to variance inflation. The population growth rate still has a significant association instead.

Now, there are two ways we can refine spatial regression: - Try to remove the spatial structure from covariates, at least from `UIS` and `ELI`, e.g. by leveraging on the spectral properties of the underlying graph. To do so, we rely on the innovative methodology of (Urdangarin et al. 2024).

Here, we do **not** mean to affect `TEP_th_22`, but only the other three covariates; this because the spatial pattern in a distance indicator should not, in my view, be removed from a regression model

- Drop those two covariates directly and only keep `TEP_th_22` and `ELI`

We show the `R` code for the first alternative approach, i.e. removing spatial trends from covariates. Considering we have 256 sites, we remove the last 13 non-constant eigenvectors (5% of the total)

```
deconfound <- function(X, Lapl = Lapl_con, n.eigen.out){
  X <- as.matrix(X)
  V <- eigen(Lapl)$vectors
  rk <- Matrix::rankMatrix(Lapl)
  coef <- solve(V, X)
  eigen.in <- c(1:(rk-n.eigen.out), c((rk+1):ncol(V)))
  X_nosp <- V[, eigen.in] %*% coef[eigen.in, ]
  return(X_nosp)
}
```

Estimated covariate effects are shown

```
##                mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -7.603 0.065     -7.734   -7.603     -7.478 -7.603   0
## TEP_th_22   -0.376 0.078     -0.529   -0.376     -0.222 -0.376   0
## PGR          0.113 0.072     -0.028    0.113      0.254  0.113   0
## UIS         -0.056 0.069     -0.191   -0.057      0.080 -0.057   0
## ELI         -0.041 0.066     -0.170   -0.041      0.090 -0.041   0
```

As we see, removing the spatial patterns from covariates does not "move" credibility interval of`UIS` and`ELI` from zero. In this particular application, we deem deconfounding covariates of scarce interest.

If, instead, we remove the non-significant covariates, this would be the result:

```
##                mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -7.608 0.065     -7.738   -7.607     -7.483 -7.607   0
## TEP_th_22   -0.354 0.072     -0.495   -0.354     -0.213 -0.355   0
## PGR          0.122 0.068     -0.011    0.122      0.255  0.122   0
```

And an improved WAIC. Please notice the decrease in WAIC is higher than the decrease in free parameters, hence the improvement must also be attributed to improved fitting.

```
cav_bym_INLA_2covs$waic$waic
```

```
## [1] 947.2792
```

```
cav_bym_INLA_2covs$waic$p.eff
```

```
## [1] 75.71434
```

This model may bring to interesting results. We can also notice a change in the mixing parameter:

```
# Model with four covariates ("full" one)
round(cav_bym_INLA$summary.hyperpar,3)
```

```
##                      mean    sd 0.025quant 0.5quant 0.975quant  mode
## Precision for ID 3.659 1.225      1.908    3.446      6.668 3.040
## Phi for ID       0.295 0.220      0.016    0.243      0.792 0.039
```

```
# Model with two covariates ("reduced" one)
round(cav_bym_INLA_2covs$summary.hyperpar,3)
```

```
##                      mean    sd 0.025quant 0.5quant 0.975quant  mode
## Precision for ID 3.219 1.078      1.647    3.041      5.838 2.707
## Phi for ID       0.395 0.225      0.050    0.371      0.846 0.203
```

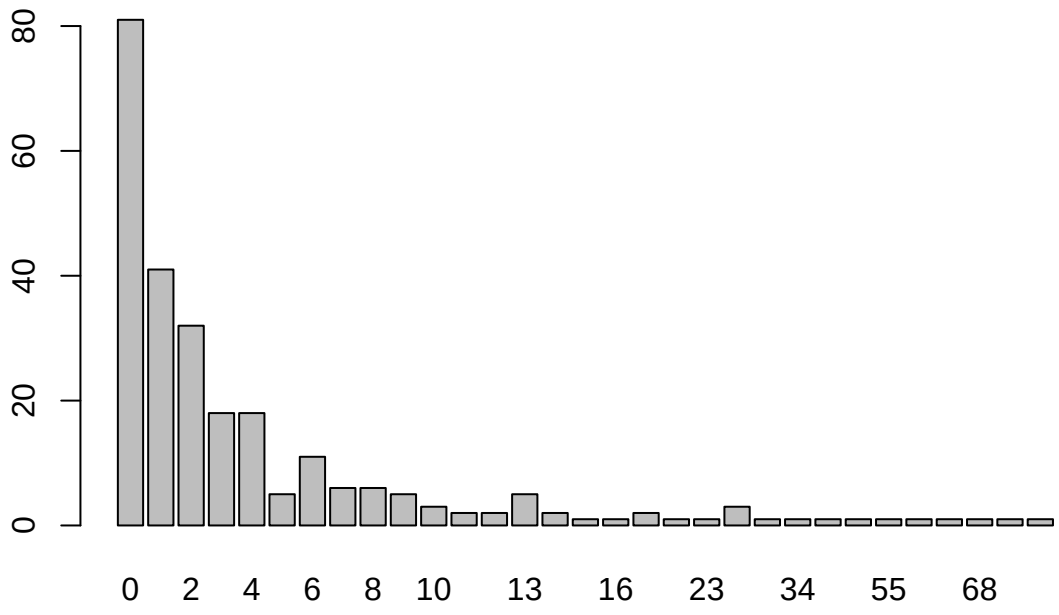
Further developments: zero-inflated regression

The number of zero counts is high:

```
sum(dd_con$N_ACC == 0)
```

```
## [1] 81
```

```
barplot(table(dd_con$N_ACC))
```



We may wonder if the data generating process incorporates a zero-generating component. We can model this augmented process through zero-inflated Poisson likelihood:

$$p(y_i|\eta_i) = \pi_0 \mathbb{I}\{y_i = 0\} + (1 - \pi_0)\mathbb{I}\{y_i > 0\} \frac{e^{-e^{\eta_i - P_i} - \eta_i y_i}}{y_i!}$$

Where $\pi_0 := \text{Prob}\{y_i = 0\}$ for all i. The remainder of the notation is akin to the nonspatial regression section.

We estimate the ZIP model maintaining the BYM latent effect inside η:

And show the corresponding summaries:

```
summary(cav_bym_zip_INLA)
```

```
## Time used:
##     Pre = 19.2, Running = 1.75, Post = 0.109, Total = 21.1
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -7.495 0.062     -7.620   -7.494     -7.377 -7.494   0
## TEP_th_22   -0.412 0.067     -0.543   -0.412     -0.281 -0.412   0
## PGR          0.103 0.064     -0.022    0.103      0.228  0.103   0
## UIS         -0.079 0.054     -0.184   -0.079      0.027 -0.079   0
## ELI         -0.058 0.055     -0.165   -0.058      0.050 -0.058   0
##
## Random effects:
##   Name     Model
##     ID BYM2 model
##
## Model hyperparameters:
##                                                          mean    sd 0.025quant
## zero-probability parameter for zero-inflated poisson_1 0.054 0.023      0.021
## Precision for ID                                       6.668 1.986      3.693
## Phi for ID                                             0.124 0.130      0.006
##                                                        0.5quant 0.975quant
## zero-probability parameter for zero-inflated poisson_1    0.051      0.109
## Precision for ID                                          6.362     11.437
## Phi for ID                                                0.079      0.496
##                                                         mode
## zero-probability parameter for zero-inflated poisson_1 0.044
## Precision for ID                                       5.768
## Phi for ID                                             0.015
##
## Watanabe-Akaike information criterion (WAIC) ...: 958.86
## Effective number of parameters .................: 61.79
##
## Marginal log-Likelihood:  -455.61
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Estimates for π_0 are low, with mean and median ≈ 0.05. The mixing parameter is highly shrunk with respect to the non-inflated regression model. The latent component in η is thus dominated by random noise.

It is, however, worth noticing how INLA manages to meet the regularity conditions to approximate the CPO more often than in the non-inflated model. Still, CPO must be re-computed manually before employing it as an evaluation metrics:

```r
sum(cav_bym_INLA$cpo$failure)
```

```
## [1] 97.93504
```

```r
sum(cav_bym_INLA$cpo$failure > 0)
```

```
## [1] 222
```

```r
sum(cav_bym_zip_INLA$cpo$failure)
```

```
## [1] 4.912799
```

```r
sum(cav_bym_zip_INLA$cpo$failure > 0)
```

```
## [1] 93
```

The reason for failures in meeting the regularity conditions to compute the CPO within INLA would deserve more attention. A preliminary guess would be that the likelihood is strongly skewed with a small sample size. Perhaps.

Weakness elements and possible developments

From this preliminary analysis, inference on spatial models is hindered by the dominance of random noise over structured spatial effects. This can be argued from the posterior distribution of the mixing parameter in the BYM model, other than from the low spatial autocorrelation parameter in the PCAR.

This means that only to a small extent the variation in y not explained by covariates can be explained by spatial structure.

On the other hand, it is difficult to assert *all* variation not explained by covariates is pure noise, otherwise we would have evidence for the lack of autocorrelation in residuals. We tested the hypothesis of no autocorrelation in GLM residuals by the Moran's I test, but in doing so we had to only test the residuals of areas with nonzero counts.

Moreover, spatial models are estimated using the INLA. While this is a broadly employed approach in epidemiology and in disease mapping, so far we did not assess how accurate the Laplace approximation has been.

To do so, we should e.g. rerun the same models using MCMC methods, e.g. using `R` libraries such as `CARBayes`, and replicating the same prior structure used.

Lastly, we did *not* model the rate at which gender violence occurs, but the occurrence of violence reports. Higher occurrence of violence reports from a given territory may thus depend on two factors: either the higher occurrence of violence in that territory, or the ease in reporting violence for the residents.

Whereas the easiest interpretation is that violence occurrence is underestimated in low-reporting areas, at the time being nothing prevents us from suspecting that the placement of support centers is at least partially strategic, i.e. the distribution of supporting centers is more dense in areas in which violence occurs, for some reason we don't know, aa a higher frequence.

Appendix: the WAIC

Following **?**, the WAIC is given by the sum of two components:

$$WAIC := 2\sum_{i=1}^{n} \mathrm{VAR}[\ln p(y_i|\theta)] - 2\sum_{j=1}^{n} \ln \mathbb{E}[p(y_i|\theta)]$$

Where θ is the full set of model parameters; the variance and the average are computed by integrating over the posterior of θ. The first addendum denotes the number of free parameters, while the second term is a measure for goodness of fit.

Appendix: R code to implement the PCAR model in `INLA`

Although it is not readily implemented in `R-INLA` (the `"besagproper"` effect is actually the Leroux model) we may base the R code on the 'INLAMSM" package (Palmí-Perales, Gómez-Rubio, and Martinez-Beneito 2021):

```
inla.rgeneric.PCAR.model <-
  function (cmd = c("graph", "Q", "mu", "initial", "log.norm.const",
                                                "log.prior", "quit"), theta = NULL) {
  interpret.theta <- function() {
    alpha <- 1/(1 + exp(-theta[1L])) # alpha modelled in logit scale
    mprec <- sapply(theta[2L], function(x) {
      exp(x)
    })
    PREC <- mprec
    return(list(alpha = alpha, mprec = mprec, PREC = PREC))
  }
  graph <- function() {
    G <- Matrix::Diagonal(nrow(W), 1) + W
    return(G)
  }
  Q <- function() {
    param <- interpret.theta()
    Q <- param$PREC *
      (Matrix::Diagonal(nrow(W), apply(W, 1, sum)) - param$alpha * W)
    return(Q)
  }
  mu <- function() {
    return(numeric(0))
  }
  log.norm.const <- function() {
    val <- numeric(0)
    return(val)
  }
  log.prior <- function() {
    param <- interpret.theta()
    val <- -theta[1L] - 2 * log(1 + exp(-theta[1L]))
    # # PC prior
    val <- val + log(lambda/2) - theta[2L]/2 - (lambda * exp(-theta[2L]/2))
    # # Gamma(1, 5e-5), default prior:
    #val <- val + dgamma(exp(theta[2L]), shape = 1, rate = 5e-5, log = T) + theta[2L]
    # # Uniform prior on the standard deviation
    #val <- val - sum(theta[2L])/2 - k * log(2)
    return(val)
  }
  initial <- function() {
    return(c(0, 4))
  }
  quit <- function() {
    return(invisible())
  }
  if (as.integer(R.version$major) > 3) {
    if (!length(theta))
      theta = initial()
  }
```

```
  else {
    if (is.null(theta)) {
      theta <- initial()
    }
  }
  val <- do.call(match.arg(cmd), args = list())
  return(val)
  }

PCAR.model <- function(...) INLA::inla.rgeneric.define(inla.rgeneric.PCAR.model, ...)
```

Bibliography

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.

Palmí-Perales, Francisco, Virgilio Gómez-Rubio, and Miguel A. Martinez-Beneito. 2021. "Bayesian Multivariate Spatial Models for Lattice Data with INLA." *Journal of Statistical Software* 98 (2): 1–29. https://doi.org/10.18637/jss.v098.i02.

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.

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.