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 during 2022. `R` codes to generate the dataset are in the R script posted here which this report is based on. Observational period are years 2021, 2022, 2023.

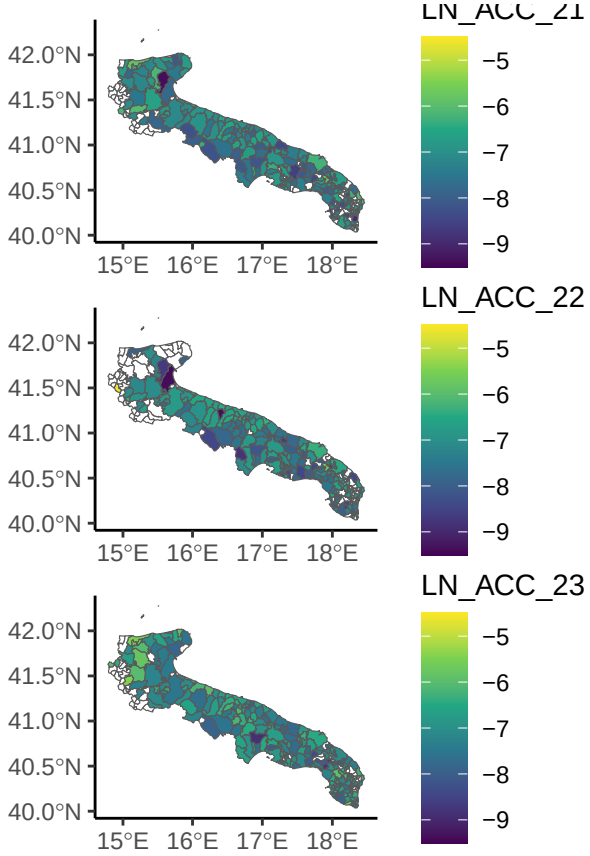
Here, we only take into account the violence reports which support centers actually take charge of, at the risk of underestimating the counts of gender-based violence cases. This choice is driven by the need of avoiding duplicated records, since e.g. it may happen that a support center redirects a victim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

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 1477, 1516 and 1822 for the three 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. Unfortunately, these data are only available for year 2021. y is modelled with simple Poisson regression.

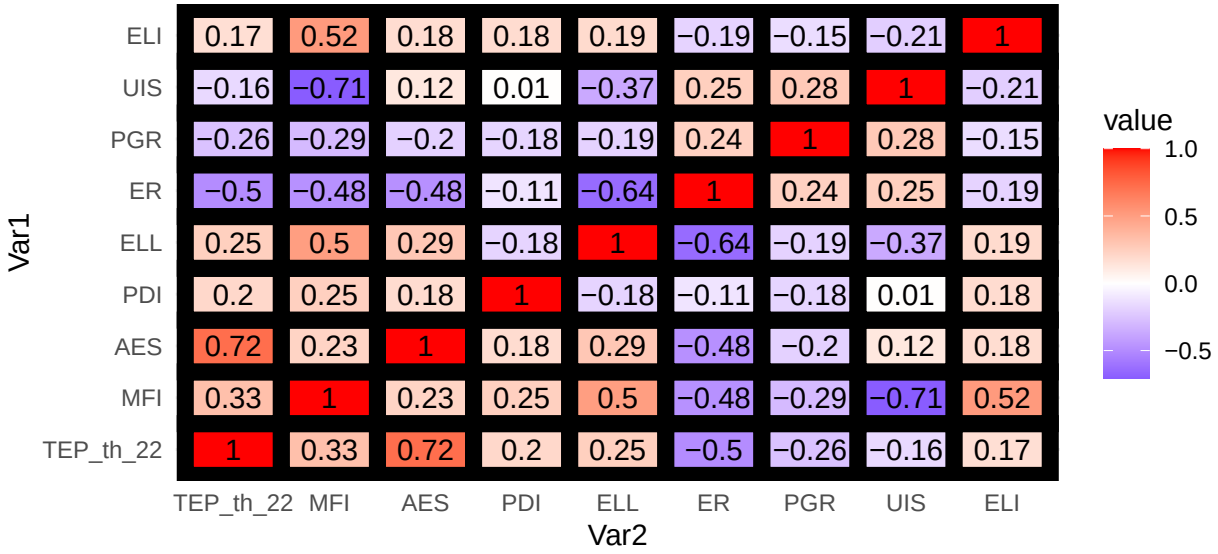
We have at disposal a number of candidate explanatory variables, which include the distance of a municipality from the closest support center and a set of variables measuring social vulnerability under different dimensions; these latter covariates are provided by the ISTAT.A more detailed description 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). 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.
- 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:



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 estimate regression coefficients, all covariates are scaled to zero mean an unit variance.

$$y_i \mid \eta_i \sim \text{Poisson}(e^{\eta_i + P_i}) \quad \text{where} \quad \eta_i = X_i^\top \beta \tag{1}$$

Where X are the covariate defined earlier, β are covariate effects, and P_i is the female population aged ≥ 15 in municipality i.

To gain more insight on the role of all explanatory variables we show the posterior summaries of the full regression model

```
cav_glm_21 <- glm(N_ACC_21~ 1 +TEP_th_22 + ELI + PGR +
                    UIS + ELL + PDI + ER,
                  family = "poisson",offset = log(nn21), data = dd_con)

cav_glm_22 <- glm(N_ACC_22~ 1 +TEP_th_22 + ELI + PGR +
                    UIS + ELL + PDI + ER,
                  family = "poisson",offset = log(nn22), data = dd_con)

cav_glm_23 <- glm(N_ACC_23~ 1 +TEP_th_23 + ELI + PGR +
```

```
                  UIS + ELL + PDI + ER,
                family = "poisson",offset = log(nn23), data = dd_con)
```

```
## # A tibble: 8 x 7
##   Effect    Mean_2021 Mean_2022 Mean_2023 Sd_2021 Sd_2022 Sd_2023
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
## 1 Intercept   -7.30     -7.47     -7.08    0.0400  0.0452  0.0367
## 2 TEP         -0.252    -0.397    -0.179   0.0366  0.0420  0.0291
## 3 ELI         -0.0371   -0.0709   -0.0491  0.0324  0.0337  0.0281
## 4 PGR          0.0357    0.129    -0.0418  0.0418  0.0427  0.0369
## 5 UIS         -0.0365   -0.0978    0.119   0.0339  0.0351  0.0295
## 6 ELL         -0.210    -0.115    -0.152   0.0440  0.0443  0.0388
## 7 PDI         -0.0492   -0.0687   -0.0681  0.0432  0.0452  0.0385
## 8 ER          -0.246    -0.0664   -0.297   0.0481  0.0502  0.0401
```

- `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 come out and 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 wuld be that in underdeveloped areas reporting gender violence is somewhat harder than in developed ones; however this relationship does not appear to be strong and is indeed negligible for 2021 and 2023 data.
- `PGR`: The association 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. Only in 2022 does growth rate appear to have a significant association with AVCs accesses.
- `UIS`: The (ventile of the distribution of the) density of production units has a somewhat ambiguous interpretation. From the one side, it has a strong negative relationship with the social frailty index. It should be therefore considered an indicator of economic development. Nevertheless, for 2022 data the regression coefficient bears the same negative sign as the incidence of low-productivity economic units; for 2023 data the association with AVCs accesses is positive instead. For 2021 data, this association appears not significantly different from zero. *Honestly I have no idea on how to interpret it.*
- `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 does not seem significantly different from zero
- `ER`: The association with employment rate is very strong and bears negative sign for 2021 and 2023 data.

Spatial modelling

We plot the log-residuals ε of the GLM regression models, defined as $\varepsilon := \ln y_i - \ln P_i - \ln \hat{y}_i$ being $\hat{y}_i$ 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 in 2022.

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.

```
spdep::moran.test(resids_glm_21[nonzero_con_21],
                  listw = spdep::nb2listw(nb_con_nonzero_21))
```

```
##
##  Moran I test under randomisation
##
## data:  resids_glm_21[nonzero_con_21]
## weights: spdep::nb2listw(nb_con_nonzero_21)
##
## Moran I statistic standard deviate = 5.6247, p-value = 9.294e-09
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.273692794      -0.005263158       0.002459683
```

```
spdep::geary.test(resids_glm_21[nonzero_con_21],
                  listw = spdep::nb2listw(nb_con_nonzero_21))
```

```
##
##  Geary C test under randomisation
##
## data:  resids_glm_21[nonzero_con_21]
## weights: spdep::nb2listw(nb_con_nonzero_21)
##
## Geary C statistic standard deviate = 4.6464, p-value = 1.689e-06
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##       0.740028075       1.000000000       0.003130524
```

```
spdep::moran.test(resids_glm_22[nonzero_con_22],
                  listw = spdep::nb2listw(nb_con_nonzero_22))
```

```
##
##  Moran I test under randomisation
##
## data:  resids_glm_22[nonzero_con_22]
## weights: spdep::nb2listw(nb_con_nonzero_22)
##
## Moran I statistic standard deviate = 5.9589, p-value = 1.269e-09
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.316941893      -0.005813953       0.002933655
```

```
spdep::geary.test(resids_glm_22[nonzero_con_22],
                  listw = spdep::nb2listw(nb_con_nonzero_22))
```

```
##
##  Geary C test under randomisation
##
## data:  resids_glm_22[nonzero_con_22]
## weights: spdep::nb2listw(nb_con_nonzero_22)
##
```

```
## Geary C statistic standard deviate = 4.3576, p-value = 6.576e-06
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##       0.731735532       1.000000000       0.003789984
```

```
spdep::moran.test(resids_glm_23[nonzero_con_23],
                  listw = spdep::nb2listw(nb_con_nonzero_23))
```

```
##
##  Moran I test under randomisation
##
## data:  resids_glm_23[nonzero_con_23]
## weights: spdep::nb2listw(nb_con_nonzero_23)
##
## Moran I statistic standard deviate = 7.6329, p-value = 1.148e-14
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.360428768      -0.005025126       0.002292370
```

```
spdep::geary.test(resids_glm_23[nonzero_con_23],
                  listw = spdep::nb2listw(nb_con_nonzero_23))
```

```
##
##  Geary C test under randomisation
##
## data:  resids_glm_23[nonzero_con_23]
## weights: spdep::nb2listw(nb_con_nonzero_23)
##
## Geary C statistic standard deviate = 5.2674, p-value = 6.917e-08
## alternative hypothesis: Expectation greater than statistic
## sample estimates:
## Geary C statistic       Expectation          Variance
##         0.7158261         1.0000000         0.0029105
```

In all these three 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.

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

$$\eta_i = X_i^\top \beta + z_i \tag{2}$$

We test a total of three models, all of which have a prior distribution depending on the spatial structure of the underlying 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 Λ as the precision parameter of the latent effect, and assign it a Wishart prior.

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

Priors for spatial effects have been defined using the `INLAMSM R` package **?**.

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:

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

Since the joint distribution of z is improper, a sum-to-zero constraint is required for identifiability.

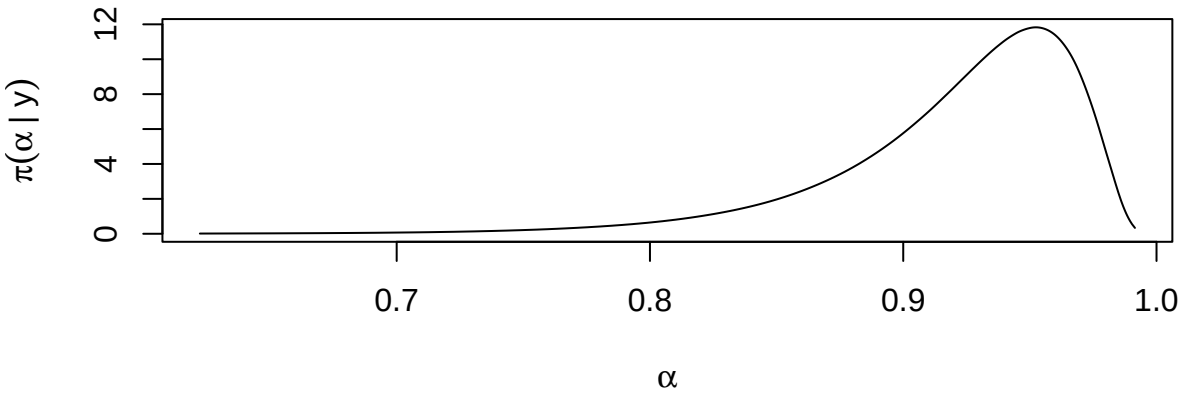
PCAR model The intrinsic autoregressive model is relatively simple to interpret and to implement, while also requiring the minimum number of additional parameter (either the scale or the precision).

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:

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

We show the posterior summary for the autocorrelation coefficient.




```
##       mean         sd quant0.025  quant0.25   quant0.5  quant0.75 quant0.975
## 0.92063945 0.04801464 0.79596553 0.89862720 0.93188514 0.95486721 0.97978353
```

The credible interval for α 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). In this case, the local prior for z_i is

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

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

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

where $R := D - W$ is the graph Laplacian matrix, W is the neighbourhood matrix and D is the corresponding degree matrix. 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, but the multivariate framework complicates the definition of the BYM model (and it does not seem to be possible to reparametrise it in such a way to have a sparse precision, hence computations are unfeasible) and hampers the use of PC priors, which would have indeed been a useful tool to prevent overfitting.

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

```
## # A tibble: 4 x 3
##   Model   WAIC Eff_params
##   <chr>  <dbl>      <dbl>
## 1 Null   3449.       76.9
## 2 ICAR   2915.      193.
## 3 PCAR   2915.      201.
## 4 Leroux 2911.      201.
```

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 Leroux model.

Since the latent effect has a proper distribution and does not require sum to zero constraints, using year-specific intercepts would cause confounding between intercepts and latent effects, leading to system crashing and thus failure to estimate the model. As a consequence, we used the same intercept for the three different years, allowing the latent effect to have *a posteriori* an additive year-specific shift.

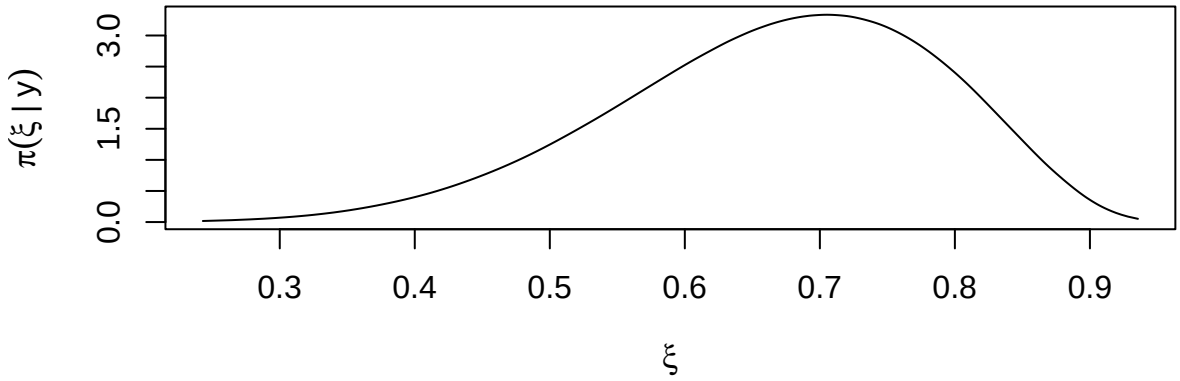
```
##      Effect Mean_2021 Sd_2021 Mean_2022 Sd_2022 Mean_2023 Sd_2023
## 1 Intercept   -7.3819  0.0864  -7.38186  0.0864   -7.3819  0.0864
## 2       TEP   -0.2452  0.0593  -0.35619  0.0728   -0.1752  0.0633
## 3       ELI    0.0158  0.0529   0.00163  0.0600   -0.0425  0.0567
## 4       PGR    0.0812  0.0617   0.13732  0.0704    0.0425  0.0631
## 5       UIS    0.0108  0.0563  -0.07407  0.0664    0.1175  0.0598
## 6       ELL   -0.2885  0.0705  -0.25615  0.0810   -0.2291  0.0744
## 7       PDI   -0.0681  0.0668  -0.02154  0.0766   -0.0777  0.0696
## 8        ER   -0.3029  0.0820  -0.15809  0.0967   -0.3285  0.0857
```

Estimations of β differ slightly from the nonspatial model. For all variables, credibility intervals are 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.

Lastly, we have a look at the mixing parameter.




```
##       mean         sd quant0.025  quant0.25   quant0.5  quant0.75 quant0.975
##  0.6672688  0.1201767  0.4078718  0.5880981  0.6777569  0.7568643  0.8682004
```

As we see, the mixing parameter is not particularly high, but interpreting it properly is hampered by the difficulty in scaling the precision matrix, which is a drawback of non-intrinsic models.

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.