Supplementary material for "Statistical Inferences and Predictions for Areal Data and Spatial Data Fusion with Hausdorff–Gaussian Processes"

November 11, 2025

S.1 An example on the Hausdorff distance

To demonstrate the efficacy of the Hausdorff distance in measuring distances between various spatial units, Figure S.1 presents three simple examples. In each sub-figure, set A is represented as a circle with a radius of 1. Figure S.1a introduces a set B, a circle tangent to A. Figure S.1b examines the distance between two sets having a non-empty intersection. Figure S.1c features B as an interior point of A. The dashed lines in the figures delineate the Hausdorff distances between the sets, highlighting the metric's ability to consider the shape, size, and orientation of the objects. This contrasts with other simpler distance functions between non-empty sets, such as the infimum distance between two sets, denoted $\inf_{x \in X} \inf_{y \in Y} d(x, y)$. This distance does not fulfill the metric property of positiveness condition. In other words, $d(x, y) = 0$ does not imply $x = y$. Indeed, employing such distance would yield a distance of zero between the three pairs of sets depicted in Figure S.1.

S.2 Complementary results for Section 3.3

In Figure S.2, we present the empirical assessment of the PEC function across four different areal datasets freely available from different `R` packages: `smile` (Godoy, 2025), `sf` (Pebesma, 2018), and `CARBayesdata` (Lee, 2022). In all of the investigated examples, we have empirical evidence the PD property holds regardless of the strength of the spatial dependence as long as $\nu \in [0, 0.8)$.

In data fusion, satellite (and computer models) data are often global and, therefore, better represented on a sphere, with distances calculated based on the Great Circle distance. Thus, we generated a synthetic map representing a hypothetical fused dataset (Figure S.3) on a sphere to empirically assess the PEC function under this important scenario. In addition, we provide an `Rcpp` script to enable fast computation of the Hausdorff distance on the sphere, since the `sf` package does not provide such a functionality. Our function takes advantage of the spatial units' convexity, a common situation in spatial data fusion. The Hausdorff distance between two convex sets equals the Hausdorff distance between their borders (Molchanov, 2005, pg. 35), thus, facilitating the implementation.
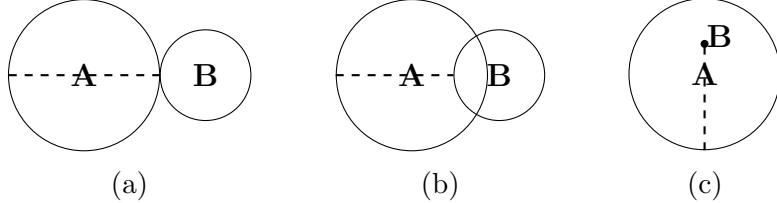
Figure S.1: Three different pairs of bounded subsets of the Euclidean plane. In (a) the sets A and B share a single point on their border. In (b), they have non-empty intersection. While in (c), B is a singleton set contained by A. The dashed lines represent the Hausdorff distance between the sets.

Figure S.4, displays our empirical assessment using the fused data (left-hand side) and only the grid-boxes on the sphere (right-hand side). The empirical evidence suggests the PEC function to be PD, except at the border of the parametric space. Although further studies are required, this result provides empirical evidence that the HGP can serve as a tool to jointly analyze global data from different satellites (or models) across different resolutions.

The empirical assessments conducted in this Section are promising. In particular, they show the HGP equipped with a PEC function may be applied to an even broader class of spatial datasets than what we initially propose in this paper, paving the way to more research on the topic in the future.

S.3 Sampling from the posterior predictive distribution of the HGP

Efficient sampling from Equation (6) utilizes the MCMC samples obtained during parameter estimation (Section 4.2). For each MCMC sample b, we generate a corresponding Monte Carlo sample from the posterior predictive distribution of $\boldsymbol{Y}^*$. First, using GP properties (Diggle et al., 1998), we sample from:

$$\big(\boldsymbol{Z}^*_{(b)} \mid \boldsymbol{Z} = \boldsymbol{z}(b)\big) \sim \mathcal{N}\big(\mu_{\boldsymbol{s}^*\mid\boldsymbol{s}}, \Sigma_{\boldsymbol{s}^*\mid\boldsymbol{s}}\big),$$

where $\boldsymbol{z}_{(b)}$ is the b-th MCMC sample of the spatial random effect, and $\mu_{\boldsymbol{s}^*\mid\boldsymbol{s}}$ and $\Sigma_{\boldsymbol{s}^*\mid\boldsymbol{s}}$ are the conditional mean and covariance matrix induced by the HGP (Equation (??)). Next, we sample $\boldsymbol{y}^*_{(b)}$ from $p(\boldsymbol{y}^* \mid \boldsymbol{\theta}_{(b)}, \boldsymbol{z}^*_{(b)})$, where $\boldsymbol{\theta}_{(b)}$ is the b-th MCMC sample of $\boldsymbol{\theta}$. This yields a Monte Carlo sample of the predictive posterior distribution of $\boldsymbol{Y}^*$, allowing estimation of various quantities associated with the distribution.

S.4 Complementary results for Section 5.2

Figure S.5 displays the map where we generate data in the areal simulation study.
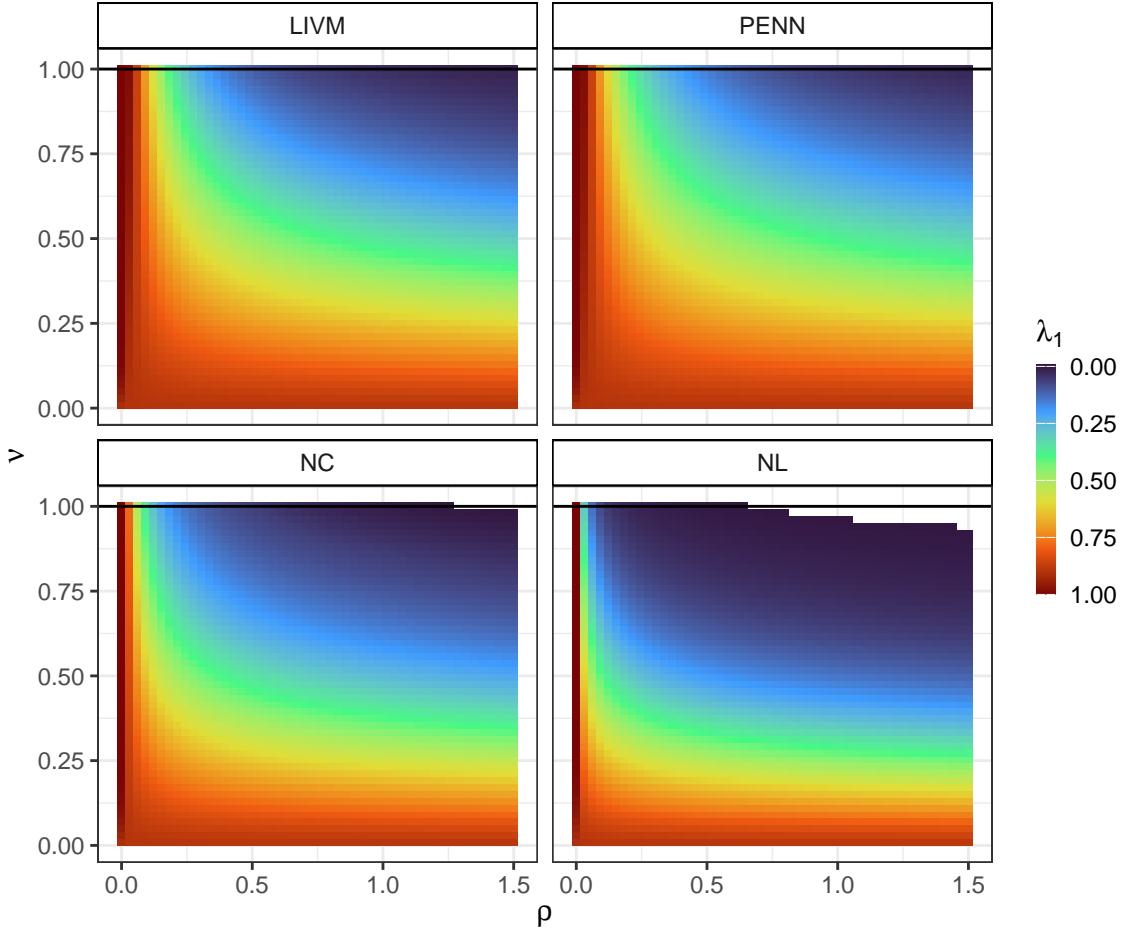
Figure S.2: Analyzing the positive definiteness of the PEC function across freely available areal datasets.

S.4.1 DAGAR priors and hyperpriors

We set $\pi(\beta_0) \propto 1$ and $\beta_1 \sim N(0, 1000^2)$ for the regression parameters; a Gamma prior with shape 2 and rate 1 was placed on the conditional precision (ϕ), and a uniform prior between 0 and 1 on the spatial dependence parameter (ψ).

S.4.2 Additional simulation results

In Table S.1, we present the bias of estimation, root mean square error (RMSE), and frequentist coverage percentage of the credible intervals for the HGP parameters under correct model specification. In terms of frequentist coverage, all parameters achieved percentages close to the nominal 95% level. The most challenging parameter, ρ, obtained a coverage of 92.5%. For estimation accuracy, measured by Mean Absolute Percentage Error (MAPE), the marginal standard deviation (σ) was a strong performer with a maximum MAPE of only 7%.
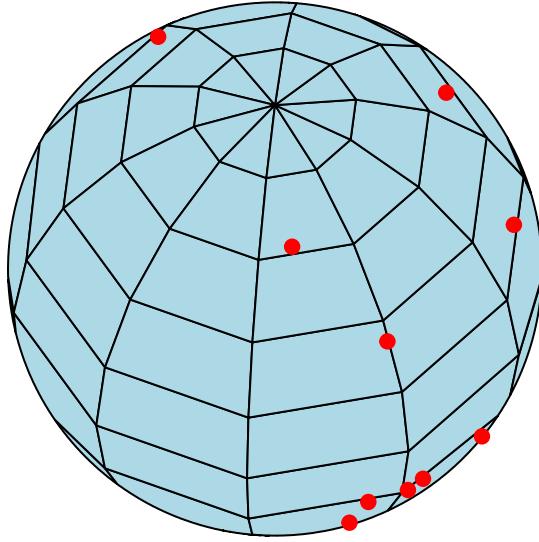
Figure S.3: Synthetically generated spatial fused data generated on a sphere.

Table S.1: Assessing parameter estimation for HGP model for areal data. Scenario indicates the data configuration scenario, True is the parameter used to simulate the datasets, Bias is the bias of estimation, MAPE is the mean absolute percentage error, RMSE is the root mean square error of estimation, and CP is the frequentist coverage percentage of the credible interval.

Scenario	Parameter	True	Bias	MAPE	RMSE	CP
	β_1	4.390	0.001	1.2	0.104	95.5
	β_2	0.330	0.001	10.9	0.063	92.5
	ρ	2.320	0.812	56.3	2.735	92.5
	σ	0.436	0.012	7.4	0.061	95.0

S.5 Complementary results for Section 5.2

S.5.1 Rescaling ρ

In our simulation study, we aimed to generate data that closely resembled the patterns observed in real-world applications. However, for the data fusion scenario, the simulated map differed from the one used in the actual application. This discrepancy necessitated a rescaling of the spatial dependence parameter, ρ, to ensure the simulated data mirrored the spatial relationships seen in the real data analyzed in Section 6.2. This adjustment was crucial because ρ is directly influenced by the scale of Hausdorff distances within a specific map.

To rescale the estimated $\rho = 13.83$, we employed a straightforward approach. First, we subtracted the smallest Hausdorff distance observed in the real data from this value. Then, we divided the resulting value by the range of Hausdorff distances present in the real data.
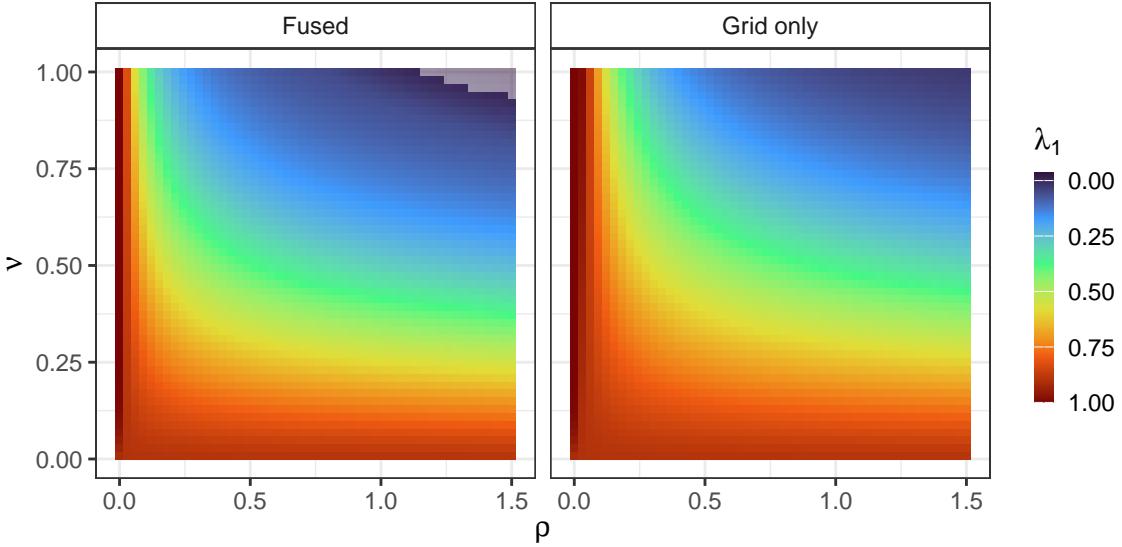
Figure S.4: Analyzing the positive definiteness of the PEC function on the sphere.

Next, we multiplied these normalized values by the range of Hausdorff distances observed in the simulated map. Finally, we added the minimum Hausdorff distance from the simulated map to each value. This process effectively yielded a spatial dependence parameter that maintained a similar degree of spatial dependence but was appropriately adjusted to the scale of the new, simulated map.

S.5.2 Generating data from an AGP

The algorithm to generate samples from the AGP is simple. We provide a detailed pseudocode to achieve that task in Algorithm 1.

Algorithm 1 Simulating data from an Aggregated Gaussian Process (AGP).

Input: A study region D, a set of n areal units $\mathcal{S} = \{s_1, \ldots, s_n\}$, a correlation function $r(\cdot; \theta)$ and its parameters θ.

Output: A vector of spatial random effects $Z = [Z(s_1), \ldots, Z(s_n)]^\top$ for the areal units.

Initialize: Generate a fine grid of $M \geq n$ points $\mathcal{G} = \{g_1, \ldots, g_M\}$ over the study region D.

Simulate a realization $[Z(g_1), \ldots, Z(g_M)]^\top$ from a zero-mean Gaussian Process with the specified correlation function $r(\cdot; \theta)$ at each grid point.

for each areal unit s_i in $\mathcal{S}$ **do**

    Identify the subset of grid points located within the areal unit: $\mathcal{G}_i = \{g_j \in \mathcal{G} \mid g_j \in s_i\}$.

    Define: $|\mathcal{G}_i|$ as the cardinality of $\mathcal{G}_i$.

    Obtain the realization for the areal unit by averaging the values at the interior grid points: $Z(s_i) \leftarrow \frac{1}{|\mathcal{G}_i|} \sum_{g_j \in \mathcal{G}_i} Z(g_j)$.

end for

return Z

Figure S.5: Study region encompassing the northern portion of Greater Glasgow, Scotland, delimited by the River Clyde. Coordinates are referenced to the OSGB36 datum and projected in the British National Grid.

Figure S.6 displays the maps used in the data fusion simulation study.

S.5.3 Inference AGP

Inference for the AGP models is carried out using the integrated nested Laplace approximation and the stochastic partial differential equation (Lindgren et al., 2011, SPDE) method. The joint distribution induced by this AGP involve approximating stochastic integrals. Practitioners must define a grid with a fixed number of points within each areal unit to approximate these integrals (Wang and Furrer, 2021). The SPDE uses a mesh to discretize the space to achieve such an approximation efficiently. The number of grid boxes per polygon will depend on the resolution of the mesh. We used two different meshes in Equation (1), which are displayed in Figure S.7.

The priors for the parameters and hyperparameters of the data fusion AGP model are taken from Moraga et al. (2017). That is, for β_0 we set a flat improper prior and a log-Gamma (shape 1 and rate 5×10^{-5}) for the log of the precision of the measurement error variance (i.e., $\log(\tau^2)$). Lastly, for $[\sigma^2, \rho]^\top$, the authors use multivariate normal prior on a non-linear transformation of these two parameters (Lindgren and Rue, 2015).

S.5.4 Additional results

Table S.2 further reveals a consistent decline in prediction accuracy for all models as the area of the spatial units decreased. For AGP-generated data, the AGP$_2$ model's RMSP
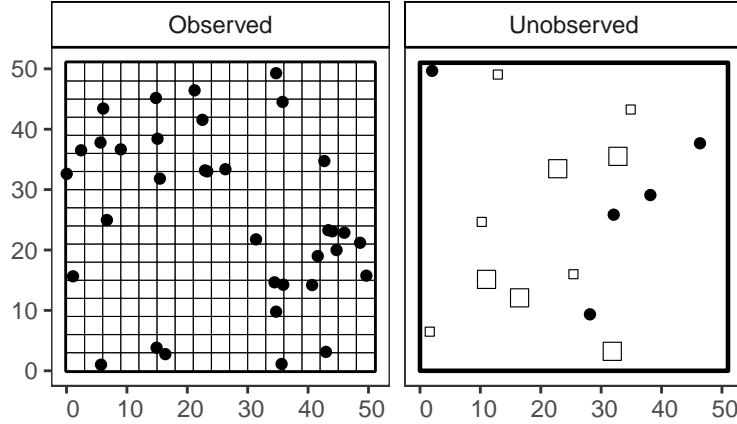
Figure S.6: Maps utilized in the data fusion simulation study. The left map displays the region used for fitting the models, while the right map shows the data points used to evaluate the models' predictive accuracy.
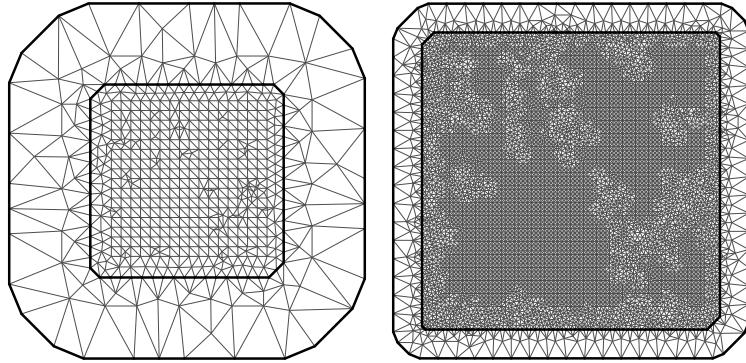

Figure S.7: This panel presents sparse and fine meshes on its top and bottom rows, respectively. The first column of the panel corresponds to the map for the simulation scenario 1, while the second column is associated to simulation scenario 2.

for point-referenced data was approximately four times higher than for the polygons with bigger area (i.e., 3^2). Similarly, the HGP model's RMSP increased by a factor of 3.15 for the same comparison (point-referenced vs largest polygon). For both the HGP and AGP_2, CPP remained near the nominal level for all the spatial units. However, HGP's CPP dropped below nominal level for smaller polygons (area of 1.5^2). This is likely due to variance underestimation: while the AGP-generated data's variance is inversely proportional to spatial units, the fitted HGP model only distinguishes between points and polygons, not their area. Regarding the IS, both our model and the AGP_2 model showed IS values roughly three times higher for point-referenced data compared to the largest polygons. This aligns with the simulation parameters. The AGP_1 model's IS performance deteriorated substantially when decreasing the area of the spatial units.

Under the HGP data-generating process, the HGP model consistently outperformed competitors on all assessment metrics. Although point predictions were broadly similar across

7

Table S.2: Out-of-sample prediction for the three models under different data generating models. "Sp. unit" indicates the geometry of the sample units for which predictions were assessed; for polygons (denoted Poly), area is additionally specified within parenthesis. Fit specifies the model fitted to the data. RMSP is the root mean squared error of prediction, CPP is the frequentist coverage percentage of the 95% prediction interval, and IS is the interval score. Double headers denote the true data-generating model.

Sp. unit	Fit	AGP			HGP		
		RMSP	CPP	IS	RMSP	CPP	IS
Poly (3^2)	AGP_1	0.78	75.8	5.79	0.89	72.5	6.96
	AGP_2	0.60	96.3	2.87	0.80	89.7	4.14
	HGP	0.83	97.9	4.00	0.74	96.4	3.36
Poly (1.5^2)	AGP_1	1.33	75.6	11.73	0.90	75.6	6.47
	AGP_2	1.13	95.8	5.38	1.09	96.6	5.32
	HGP	1.50	79.2	9.36	0.85	94.8	4.01
Points	AGP_1	2.56	52.2	35.70	3.64	22.8	80.70
	AGP_2	2.43	95.1	11.54	3.51	77.9	22.42
	HGP	2.62	95.3	12.57	2.94	96.0	13.79

models, the HGP provided remarkably superior interval predictions for all spatial unit types, mantaining CPP near the nominal level (see Table S.2). In contrast, the AGP_1 model exhibited a CPP of only 22.8% for point-referenced data. The poor coverage for both AGP models with point-referenced data is further reflected in their IS, which were 5.9 (AGP_1) and 1.6 (AGP_2) times higher than the HGP. On the other hand, the AGP_2 model remained competitive with the HGP for polygons interval predictions, while the sparse AGP performed poorly.

In Table S.3, we present the bias of estimation, root mean square error (RMSE), and frequentist coverage percentage of the credible intervals for the model parameters under the HGP correct specification. The MAPE for ρ was around 10.9% while for σ_a it was 9.21%.

Table S.3: Assessing parameter estimation for HGP model for fused data. Map indicates the size of the grid where the data was simulated, True is the parameter used to simulate the datasets, Bias is the bias of estimation, MAPE is the mean absolute percentage error, RMSE is the root mean square error of estimation, and CP is the frequentist coverage percentage of the credible interval.

Parameter	True	Bias	MAPE	RMSE	CP
β_1	5.810	−0.022	4.903	0.358	90.0
ρ	27.934	0.003	10.927	3.836	99.5
σ_a	1.311	−0.115	9.121	0.145	84.5
σ	4.442	−0.150	8.616	0.477	93.5
τ	0.170	0.170	99.788	0.181	79.5

Table S.4: Different model comparison criteria for the homoscedastic HGP models fit with different values for ν.

	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$
LOOIC	1091.9	**1081.5**	1092.6	1090.8
WAIC	1039.3	**1038.7**	1039.3	1039.2

Similarly to what is observed with geostatistical models, the small scale standard deviation parameter, namely τ, was the most challenging parameter to be estimated. Unlike the areal simulation, the coverage of some of the variance parameters was consistently below the nominal level of 95%. This results are expected, as the HGP inherits the weak identifiability of these parameters from geostatistical models under a normal likelihood (Zhang, 2004). Predictions remain trustworthy despite this issue.

S.6 Complementary results for Section 6.1

To choose the most appropriate value of ν in our application, we have fitted out model with $\nu \in \{0.6, 0.7, 0.8, 0.9\}$. These values were selected because they fill different regions of the parameter space where we have evidence of the the PEC function positive definiteness. The goodness-of-fit metrics associated to each value of ν are presented in Table S.4.

In Figure S.8, we display the observed SMR along with its smoothed estimates provided by the three different models we fit to this data. The smoothed estimates look nearly identical, further suggesting the three models deliver very similar goodness-of-fit for these data.

Figure S.9 shows the prior and empirical posterior distributions of ρ.

Figure S.10 shows the spatial correlations induced by the DAGAR and HGP in the random effects as functions of the Hausdorff distance. Once again, it is noticeable that the HGP has less uncertainty around its estimates of spatial correlations. In addition, the upper bound of the spatial correlation function of the HGP touches the mean curve of the DAGAR, while the lower bounds of the two distributions of curves are very close to each other.

S.6.1 Convergence check for areal application

Figures S.11 and S.12 display the trace and density plots for the HGP model parameters. In addition, Figures S.13 and S.14 present the same analysis for some of the sampled random effects.

Figures S.15 and S.16 display the trace and density plots for the BYM model parameters. In addition, Figures S.17 and S.18 present the same analysis for some of the sampled random effects.

Figures S.19 and S.20 display the trace and density plots for the DAGAR model parameters. In addition, Figures S.21 and S.22 present the same analysis for some of the sampled random effects.
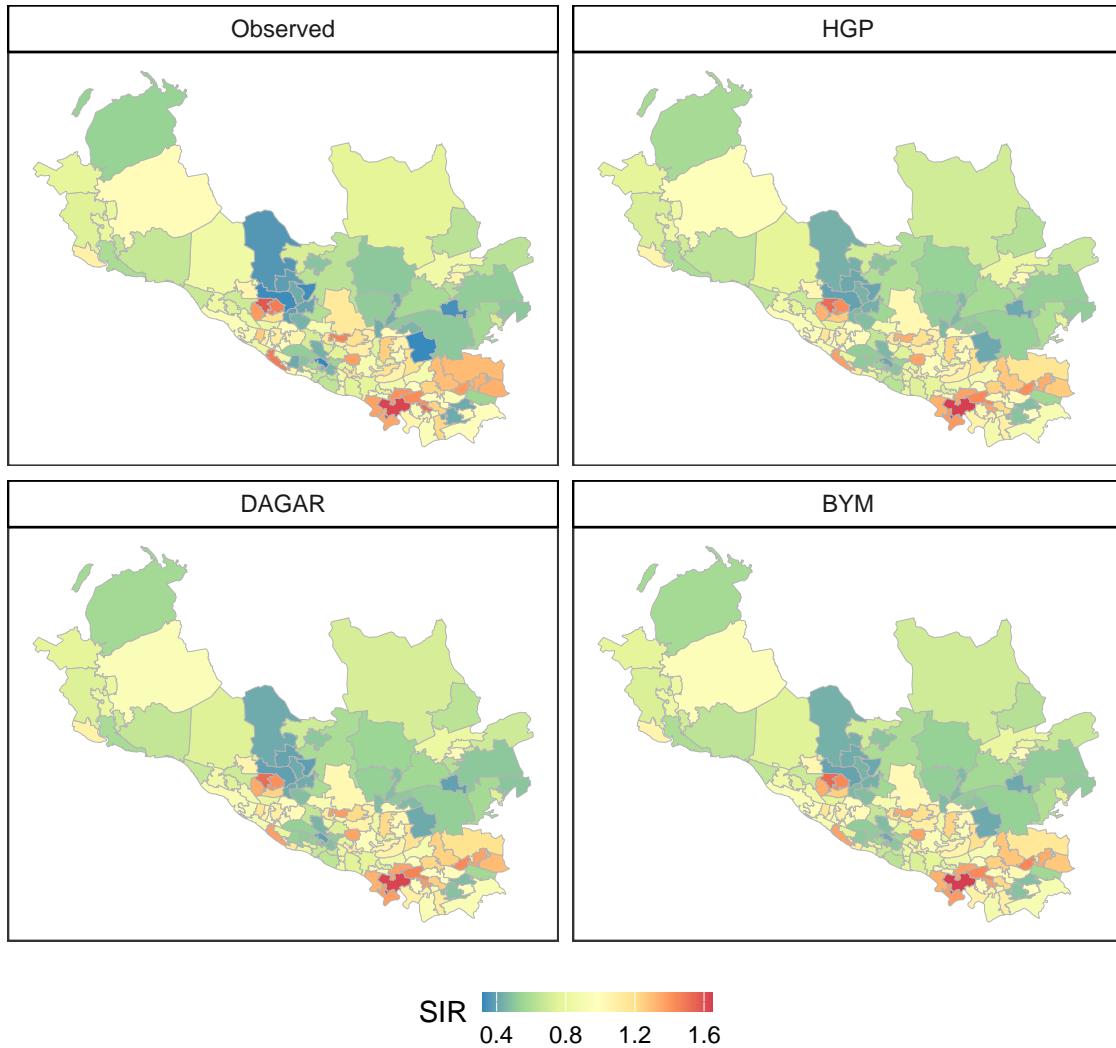
Figure S.8: Panel presenting the observed SMR and its estimated values from the HGP, DAGAR, and BYM, respectively.
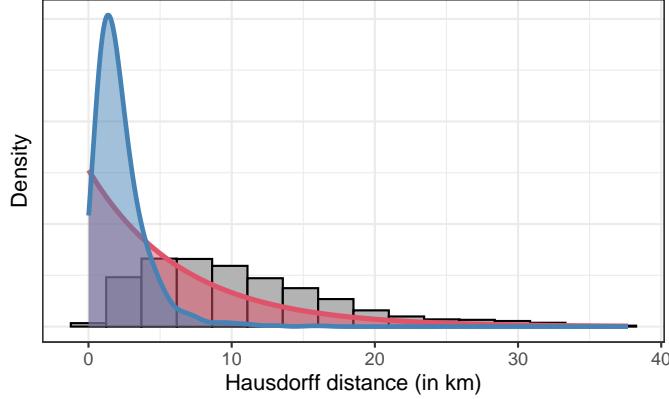
Figure S.9: Density histogram of the pairwise Hausdorff distances observed in the dataset and. In blue, estimated posterior density of the practical range ρ, and, in red, the prior used for ρ.

Table S.5: Different model comparison criteria for the heteroscedastic HGP models fit with different values for ν.

Model	Criterion	$\nu = 0.6$	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$
HGP	LOOIC	508.6	**504.5**	508.8	509.0
	WAIC	514.7	**508.2**	508.4	510.8

S.7 Complementary results for Section 6.2

Figure S.23 displays the data used in the Data Fusion application.

Figure S.24 presents the different meshes used by the AGP models in the data fusion application.

The most appropriate value of ν in our application was chosen based on the goodness-of-fit when fitting our model with $\nu \in \{0.6, 0.7, 0.8, 0.9\}$. These values were selected because they fill different regions of the parameter space where the PEC function is positive definite. The goodness-of-fit metrics associated to each value of ν are presented in Table S.5.

Table S.6 presents the predictive performance of the HGP and aggregated models using a 10-fold cross-validation. It summarizes how each model performed specifically for the point referenced and areal (polygon) data. In terms of point prediction, HGP has shown a better performance regardless of the geometry type. The same behavior was observed for interval predictions. The AGP models tend to be overconfident in their interval predictions for point-referenced data, resulting in very low coverage.

In Table S.7, we use the 10f CV to the HGP and AGP models to two alternatives. First, two GPs where the areal data are reduced to their centroids. In that case, inference and predictions are obtained as usual with a Geostatistical model. We tested an homoscedastic and a heteroscedastic "centroid GP". The latter assumes the marginal variances of the random effects vary depending on the source of the data, i.e., satellite- or measurement stations-based. Lastly, we also fit a Log-Normal model with HGP random effects. For that

11
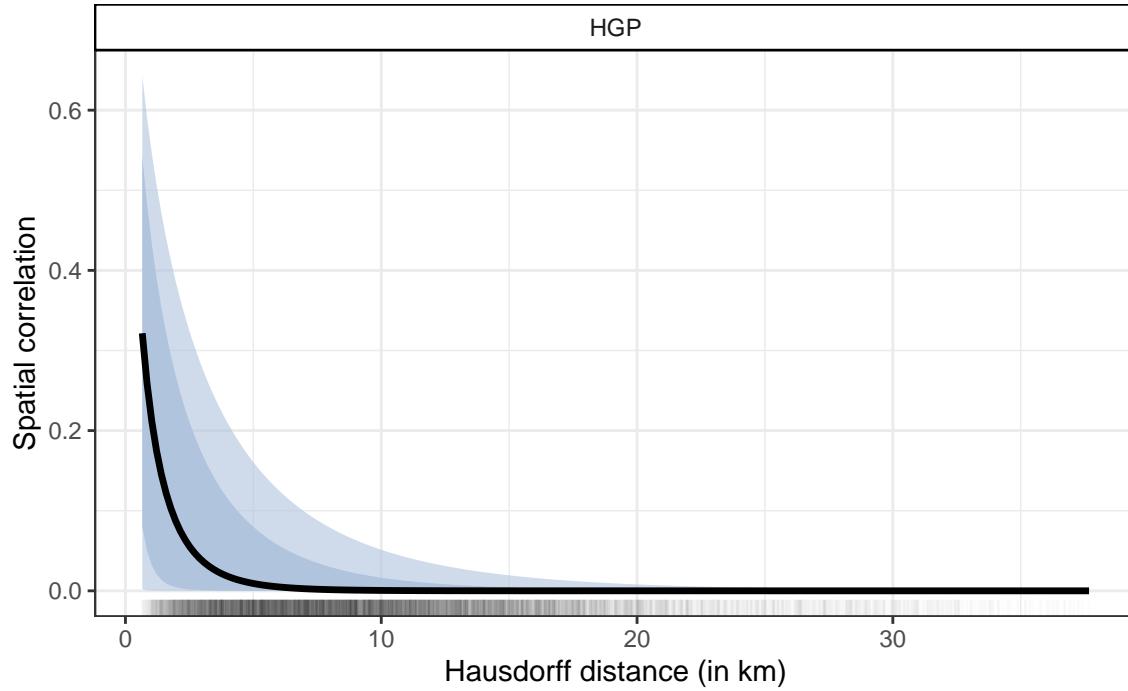
Figure S.10: Spatial correlation between random effects as a function of the Hausdorff distance.

Table S.6: Metrics for assessment of out-of-sample predictions using 10-fold cross-validation for different geometry resolutions. RMSP is the root mean squared error of prediction; CPP is the frequentist coverage of the 95% prediction interval; and IS is the interval score.

Resolution	Model	RMSP	CPP	IS
Point	HGP	2.71	85.4	12.69
	AGP_1	4.07	3.12	111.84
	AGP_2	2.71	57.3	45.03
Polygon	HGP	0.62	96.9	3.83
	AGP_1	0.86	86.6	6.03
	AGP_2	1.10	98.9	10.92
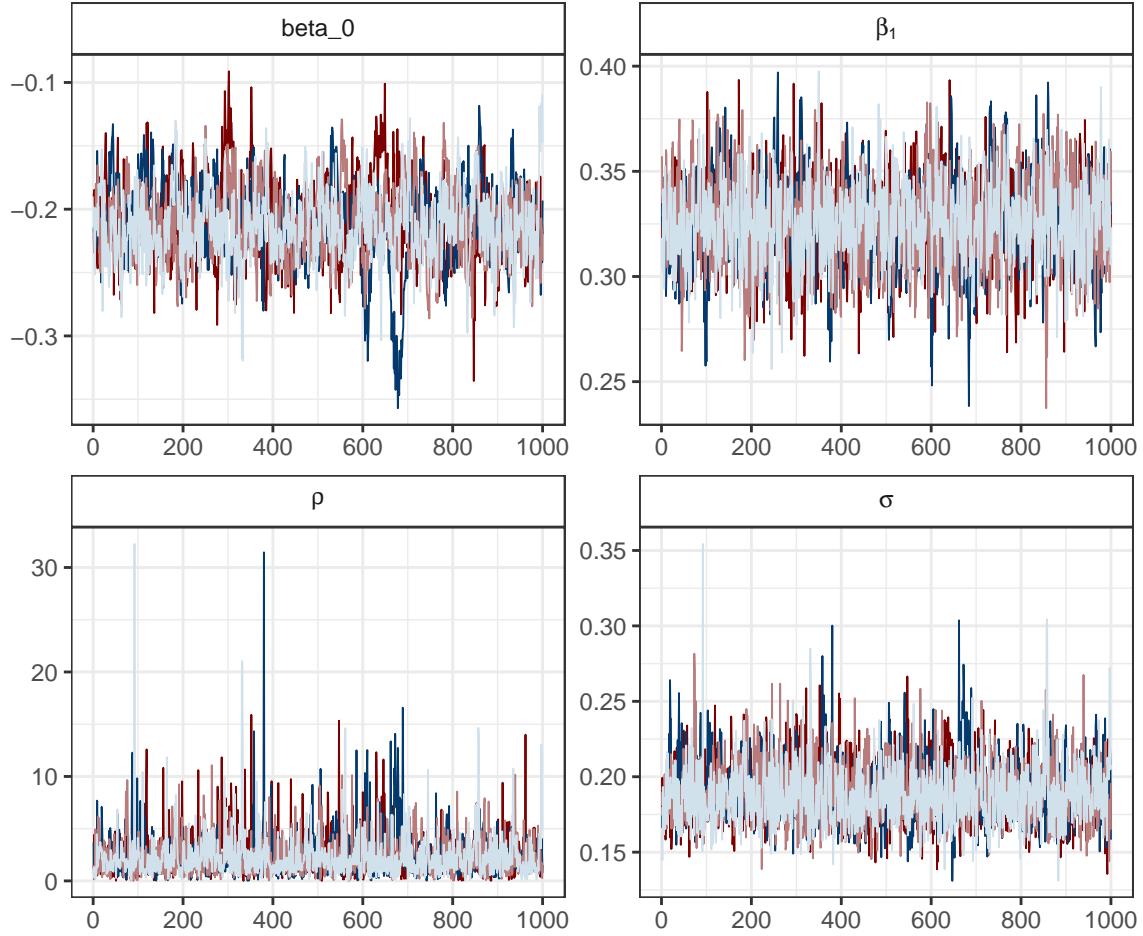
Figure S.11: Traceplots of the parameters associated to the HGP model fitted to the areal data.

model, we also considered analogous heteroscedastic and homoscedastic versions. In both cases, the covariance function was a PEC function with ν fixed at 0.7. That choice, although arbitrary, is based on that ν value consistently leading to the best fit HGPs in the examples studied here. The HGP model we used in our application performed the best, followed by the Log-Normal model with HGP random effects. The AGP models had the worst performance, indicating even the simplistic centroid-based GP is preferred in this case. However, we acknowledge that this may be related to the covariance function choice. The AGP models are based on a smoother covariance function, while the centroid-based GP uses the same covariance function employed to the HGP models. To choice for the covariance function used with the AGPs were based on the literature (Moraga et al., 2017).

S.7.1 Convergence check for data fusion application

Figures S.25 and S.26 display the trace and density plots for the HGP model parameters.
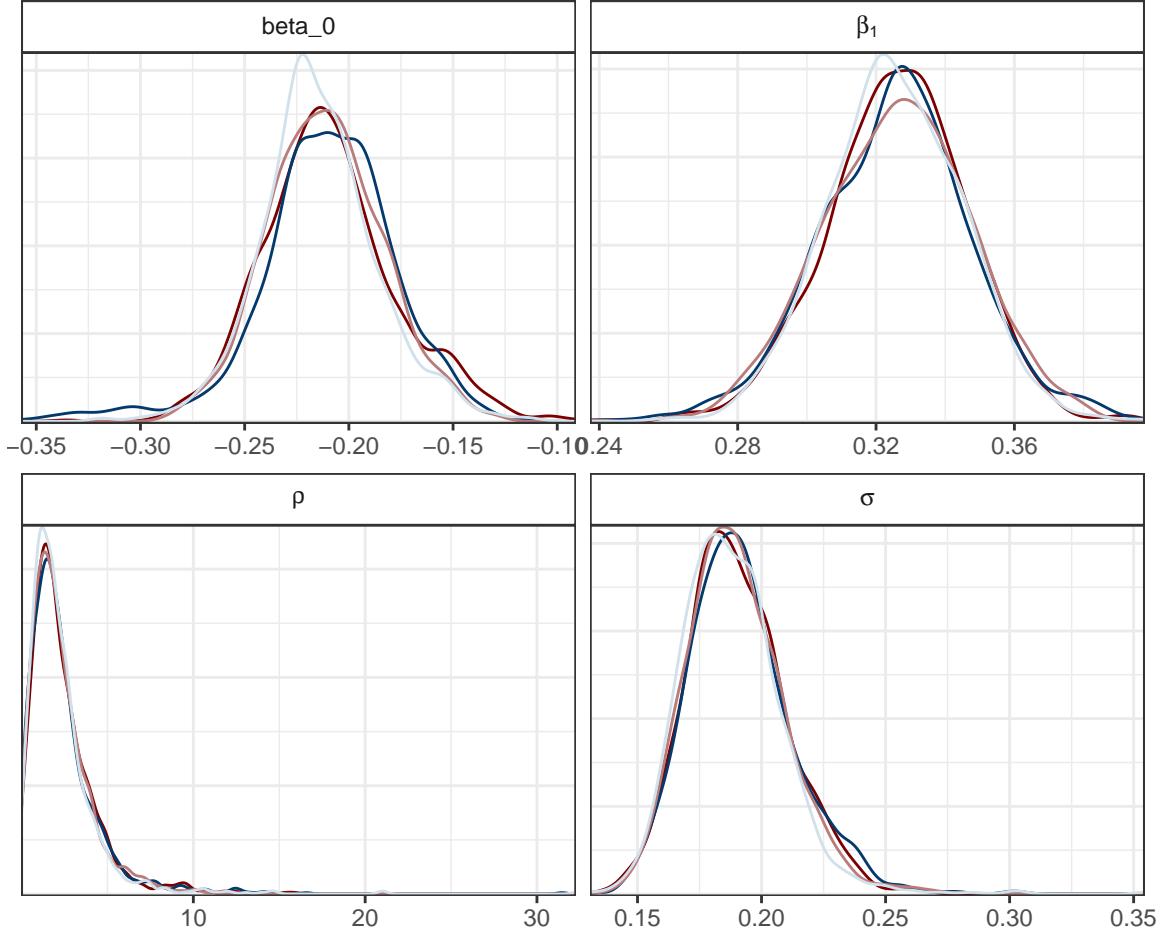
Figure S.12: Estimated densities of the parameters associated to the HGP model fitted to the areal data.

Table S.7: Metrics for assessment of out-of-sample predictions using 10-fold cross-validation for different models. RMSP is the root mean squared error of prediction; CPP is the frequentist coverage of the 95% prediction interval; and IS is the interval score. In the Model column, Gaussian-HGP is the HGP model used in Section 6.2, LN-HGP is a Log-Normal model with HGP random effects, cGP_1 and cGP_2 are homoscedastic and heteroscedastic centroid-based GPs, respectively. The AGP_1 and AGP_2 models are aggregated GPs with a sparse and fine mesh, respectively.

Model	RMSP	CPP	IS
Gaussian-HGP	**2.22**	95.6	**5.1**
LN-HGP	2.65	96.0	5.6
cGP_1	2.66	91.6	7.5
cGP_2	3.54	94.0	10.9
AGP_1	3.55	77.3	15.4
AGP_2	4.45	93.6	9.6

14
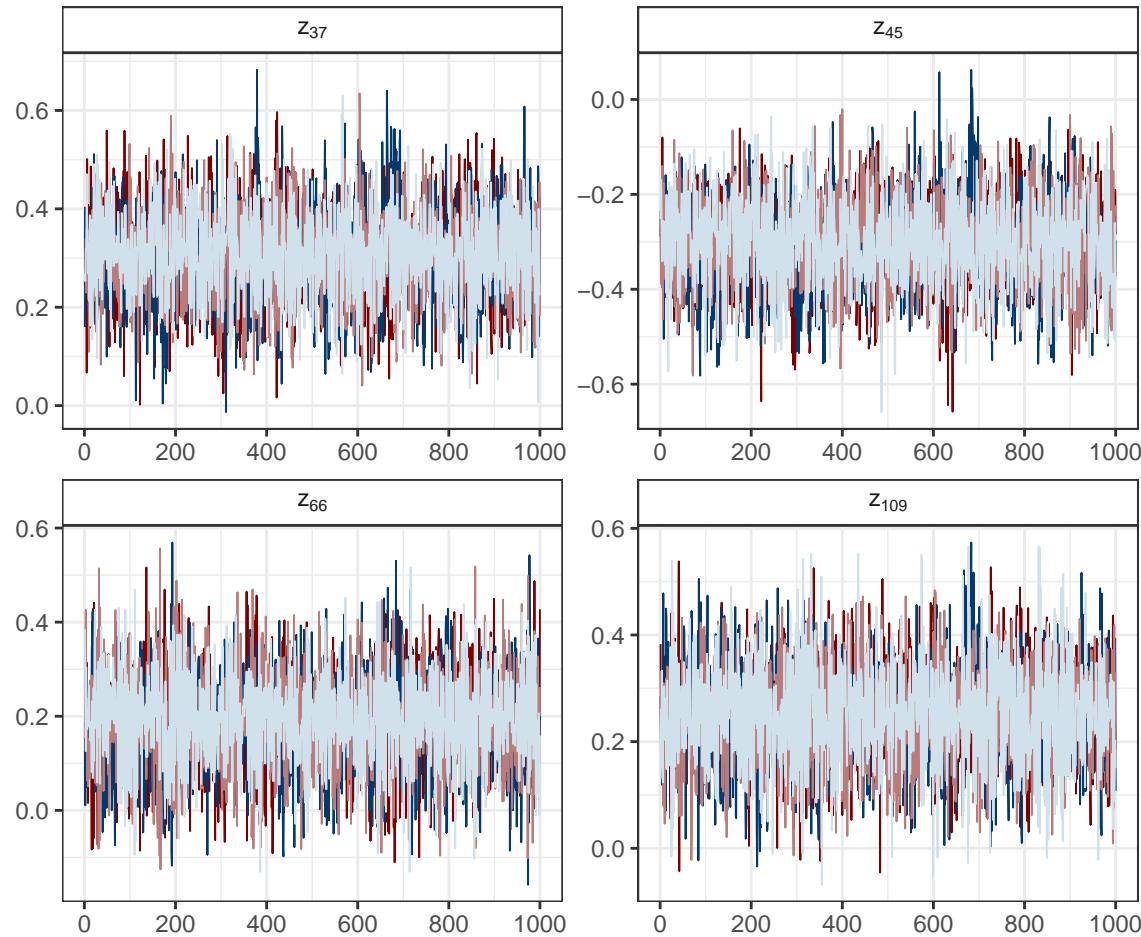
Figure S.13: Traceplots of randomly chosen random effects from the HGP model fitted to the areal data.
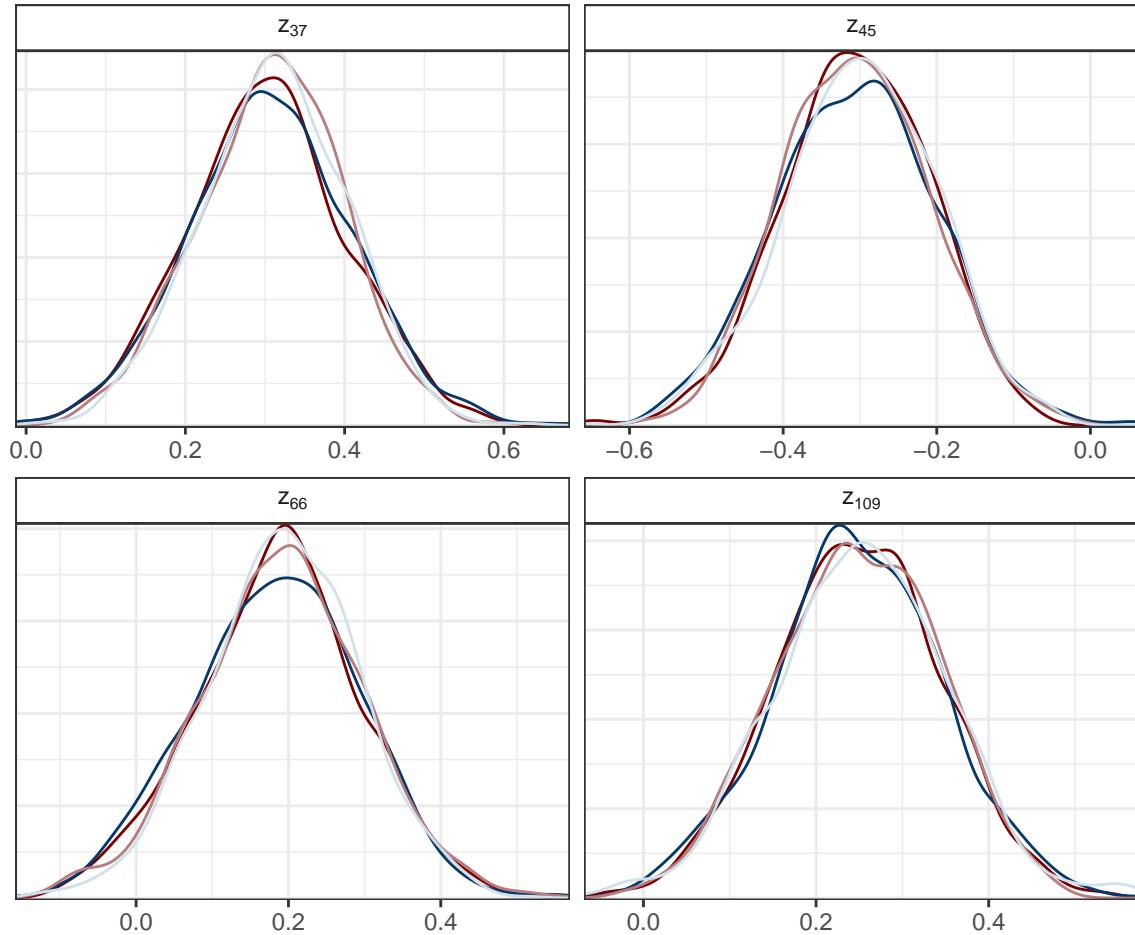
Figure S.14: Estimated densities of randomly chosen random effects from the HGP model fitted to the areal data.
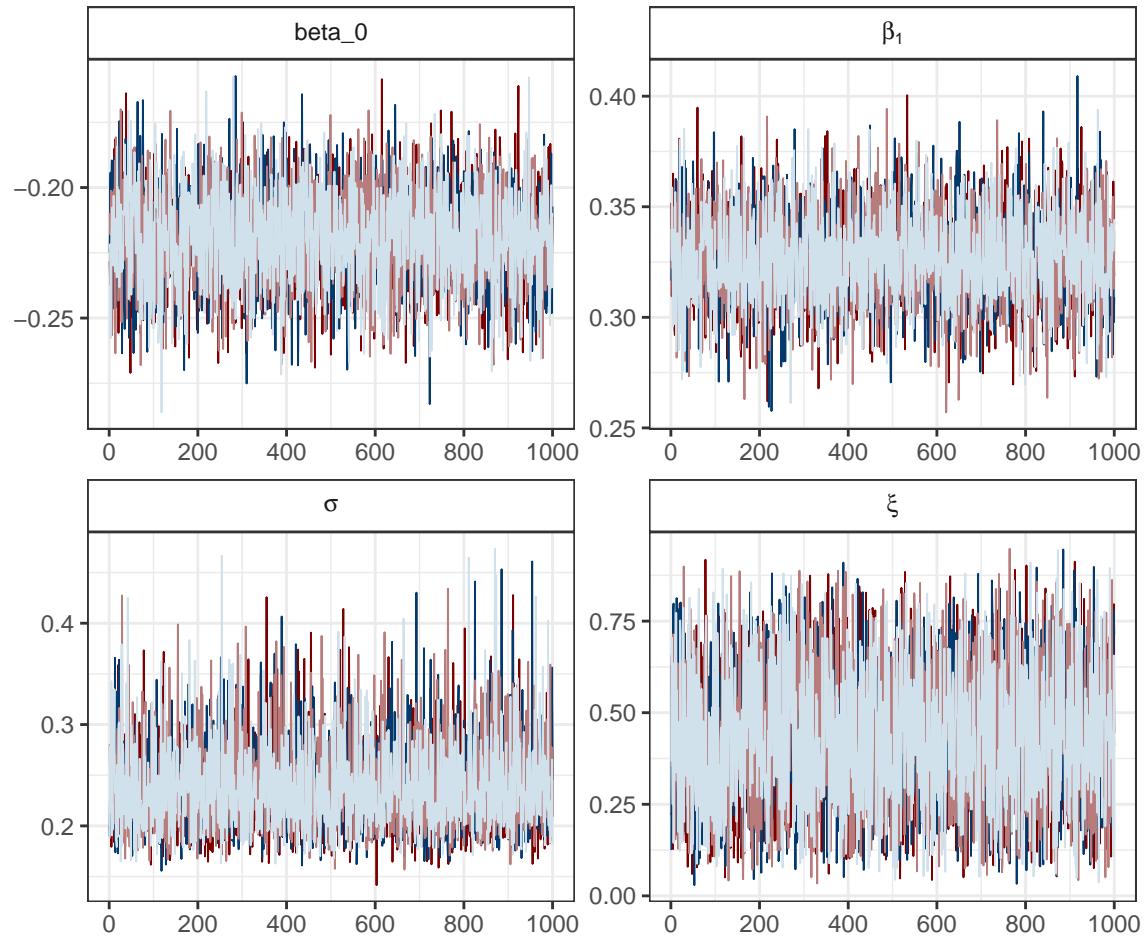
Figure S.15: Traceplots of the parameters associated to the BYM model fitted to the areal data.
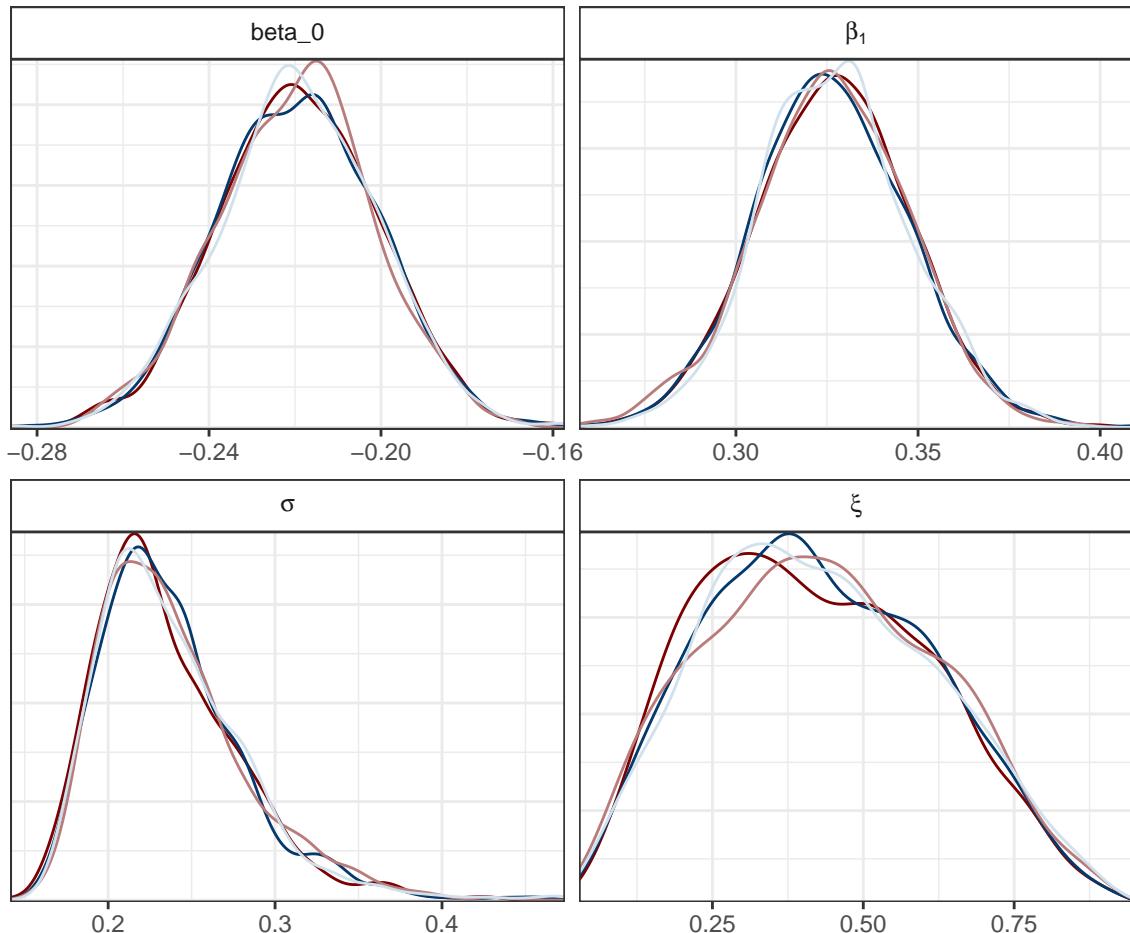
Figure S.16: Estimated densities of the parameters associated to the BYM model fitted to the areal data.
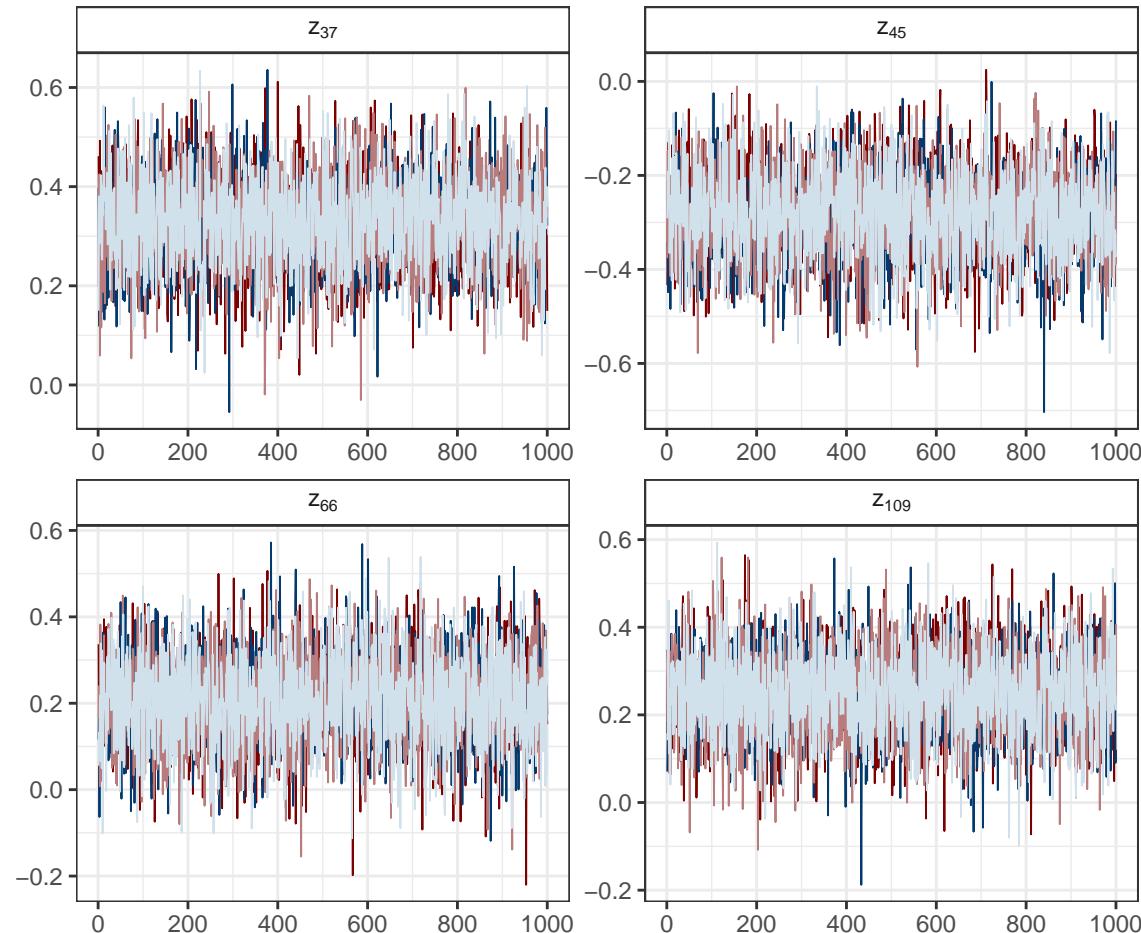
Figure S.17: Traceplots of randomly chosen random effects from the BYM model fitted to the areal data.
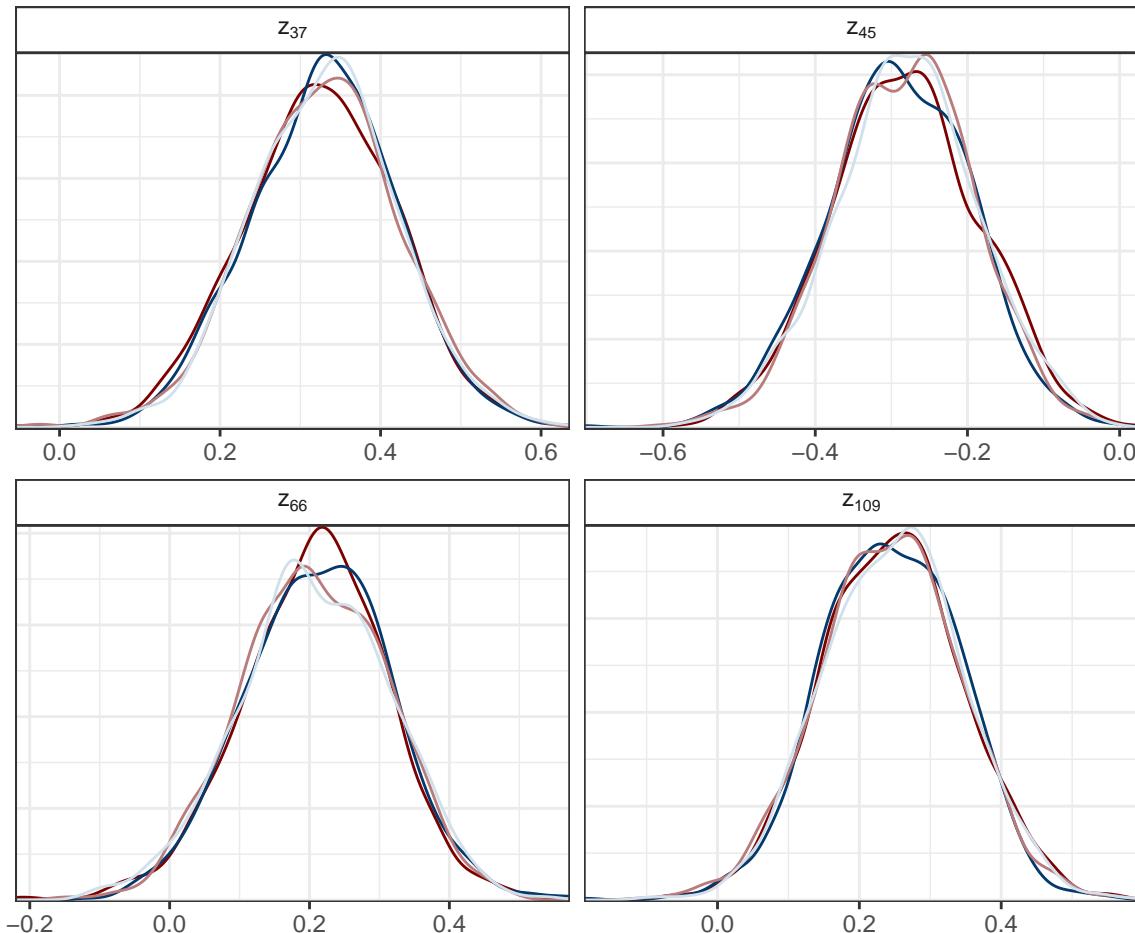
Figure S.18: Estimated densities of randomly chosen random effects from the BYM model fitted to the areal data.
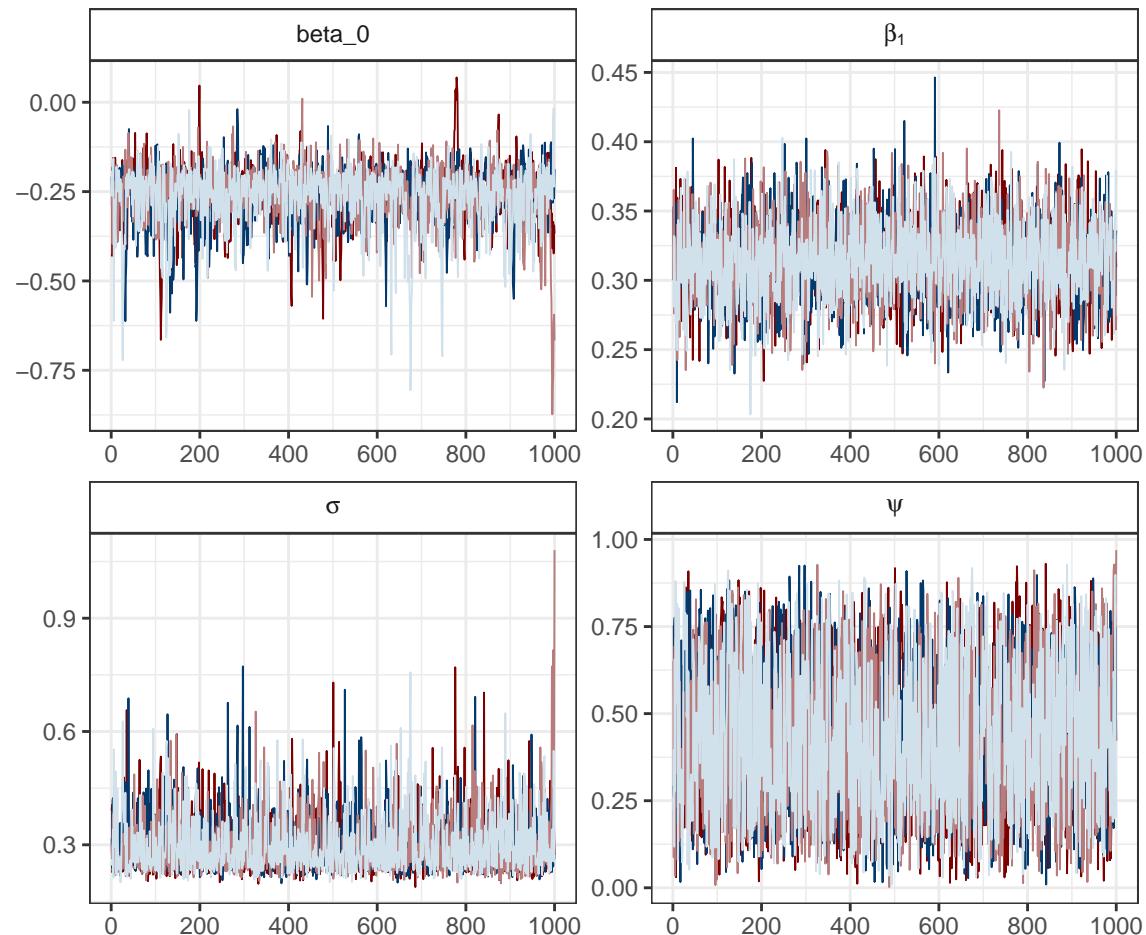
Figure S.19: Traceplots of the parameters associated to the DAGAR model fitted to the areal data.
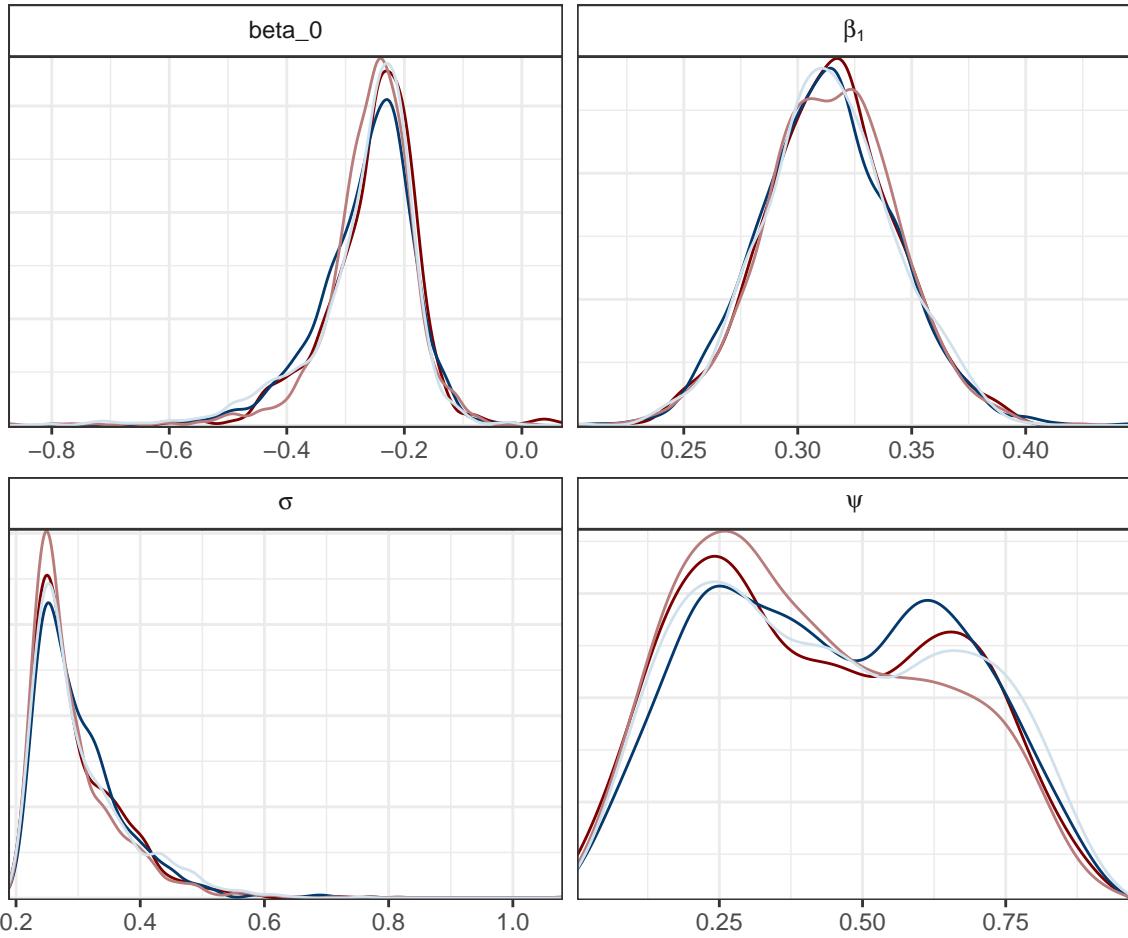
Figure S.20: Estimated densities of the parameters associated to the DAGAR model fitted to the areal data.
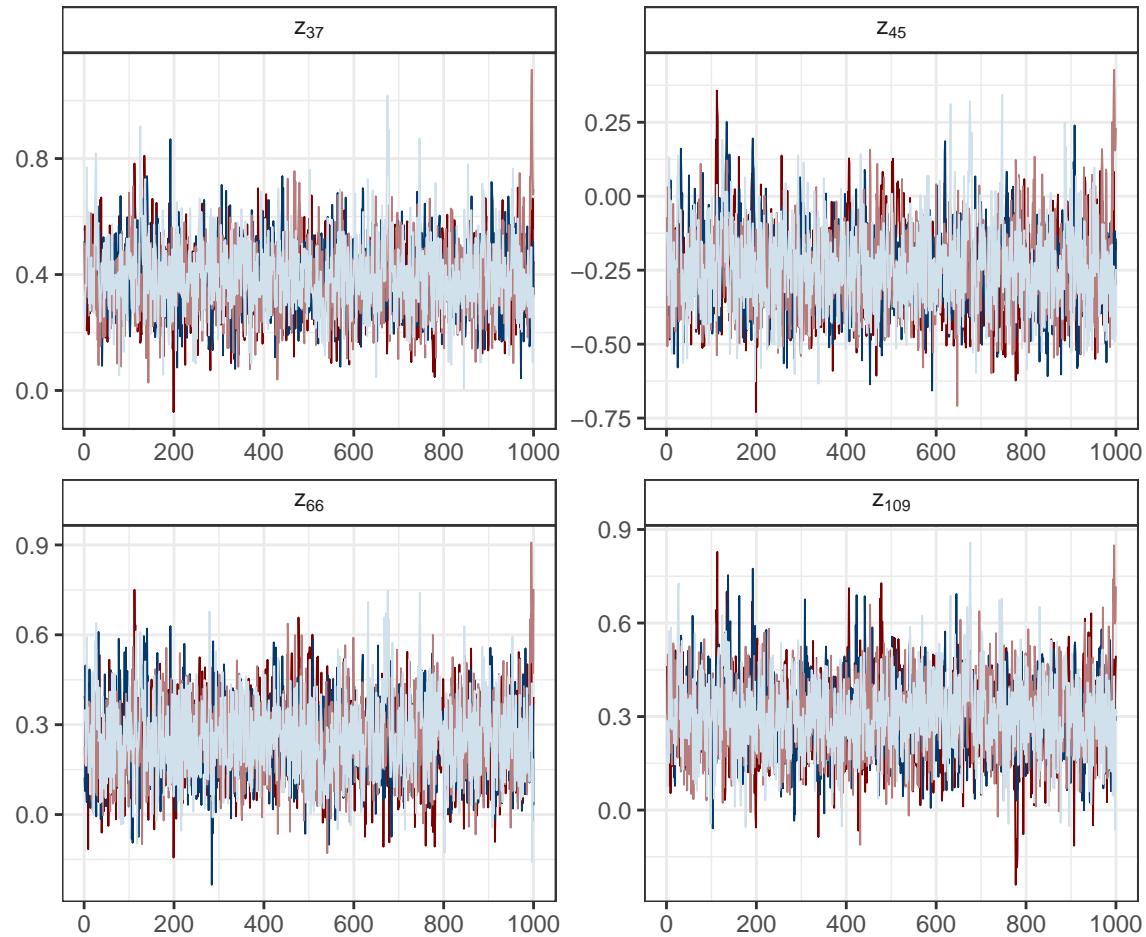
Figure S.21: Traceplots of randomly chosen random effects from the DAGAR model fitted to the areal data.
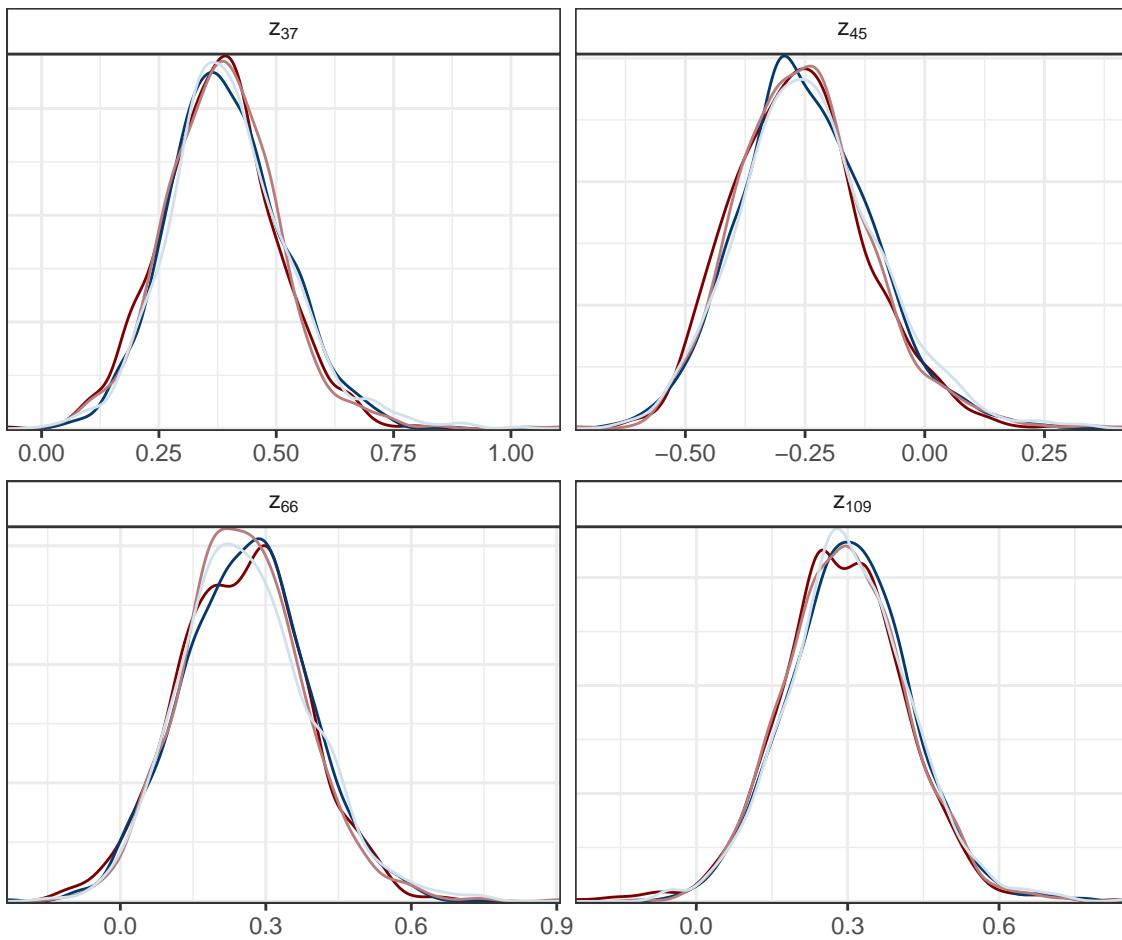
Figure S.22: Estimated densities of randomly chosen random effects from the DAGAR model fitted to the areal data.
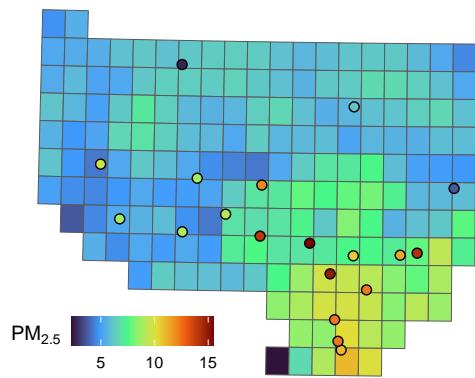

Figure S.23: PM$_{2.5}$ in Ventura and Los Angeles counties in California, United States. Points represent direct measurements. Tiles are satellite derived estimates.
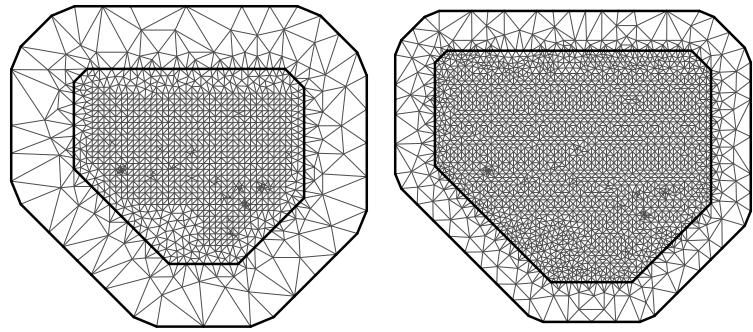
Figure S.24: The sparse and the fine meshes used in the air pollution application on the left and right side of the panel, respectively.
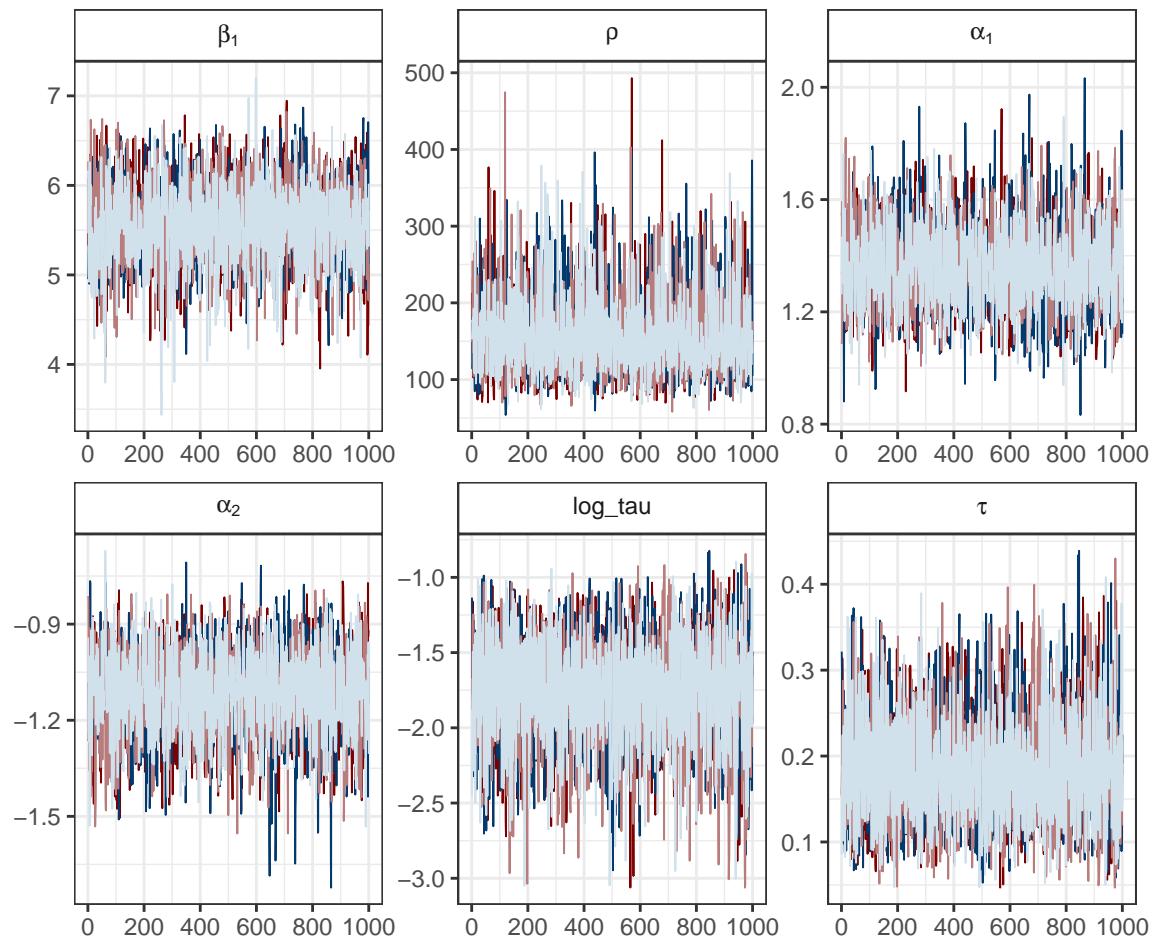

Figure S.25: Traceplots of the parameters associated to the HGP model fitted to the fused data.
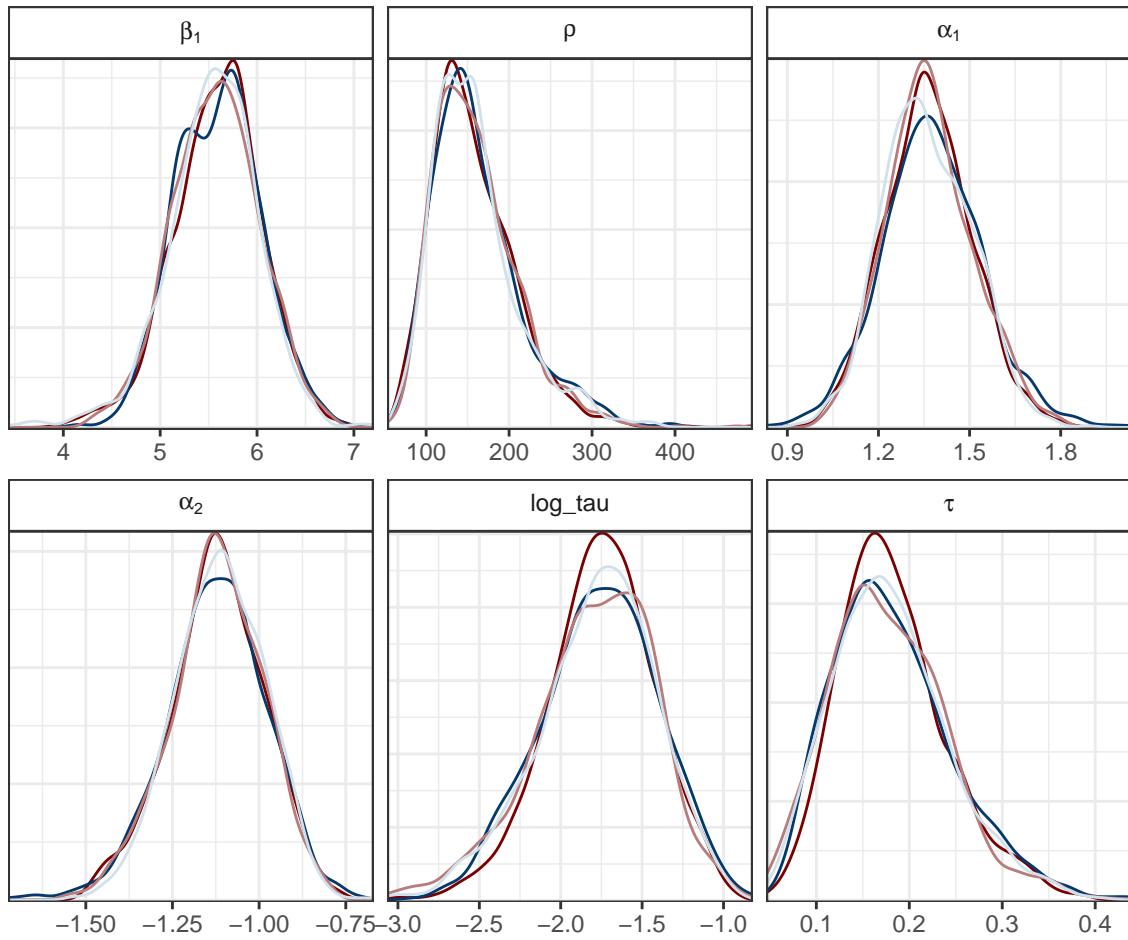
Figure S.26: Estimated densities of the parameters associated to the HGP model fitted to the fused data.

References

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47, 299–350.

Godoy, L. d. C. (2025), "smile: Spatial misalignment, interpolation, linkage, and estimation," R package version 1.1.0.

Lee, D. (2022), *CARBayesdata: Data used in the vignettes accompanying the CARBayes and CARBayesST packages*, r package version 3.0.

Lindgren, F. and Rue, H. (2015), "Bayesian spatial modelling with R-INLA," *Journal of Statistical Software*, 63.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498.

Molchanov, I. S. (2005), *Theory of random sets*, Probability and Its Applications, London, England: Springer, 2005th ed.

Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017), "A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE," *Spatial Statistics*, 21, 27–41.

Pebesma, E. (2018), "Simple features for R: Standardized support for spatial vector data," *The R Journal*, 10, 439–446.

Wang, C. and Furrer, R. (2021), "Combining heterogeneous spatial datasets with process-based spatial fusion models: A unifying framework," *Computational Statistics & Data Analysis*, 161, 107240.

Zhang, H. (2004), "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99, 250–261.