Bayesian Spatial Clustering of Disease Incidence Rates Using an Adaptive Geographically Weighted Chinese Restaurant Process

Sepideh Asadi and Mehdi Moradi

Department of Mathematics and Mathematical Statistics
Umeå University, Sweden

October 21–23, 2026 | Leiden, The Netherlands
Bayes 2026

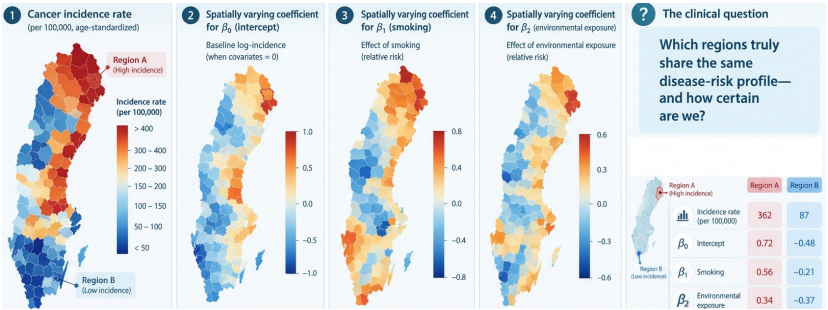


Spatial Heterogeneity in Disease Mapping

- Disease burden and risk-factor associations may vary across regions.
- Global models may obscure local heterogeneity relevant to geographically targeted public-health planning.

Spatially varying coefficient (SVC) Poisson model

$$Y_i \mid \beta(s_i) \sim \text{Poisson}(\mu_i), \qquad \mu_i = E_i \exp\left\{ x_i^{\mathsf{T}} \beta(s_i) \right\}.$$



Goal: $\beta(s_i) = \beta_{z_i}$, $z_i \in \{1, \ldots, K\}$: group regions with similar regression profiles.

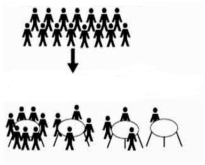
Chinese Restaurant Process: Flexible Prior, No Geography

- A Bayesian nonparametric prior over partitions.
- The number of clusters is not fixed in advance.

$$\Pr(z_i = k \mid z_{1:i-1}, \alpha) = \begin{cases} \dfrac{n_k^{(-i)}}{i-1+\alpha}, & \text{existing cluster } k, \\ \dfrac{\alpha}{i-1+\alpha}, & \text{new cluster.} \end{cases}$$

Limitation

The standard CRP allocation rule does not explicitly incorporate geographic distance.





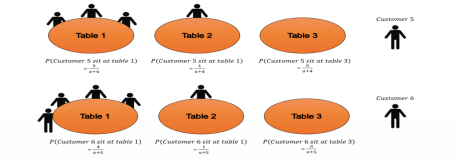


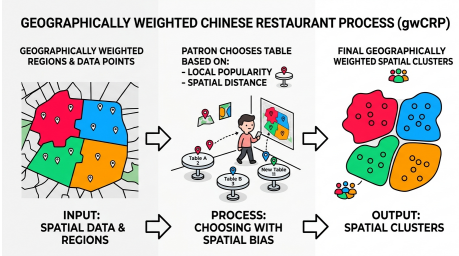
Figure: Graphical representation of the Chinese Restaurant Process (CRI

gwCRP: Incorporating Geographic Distance — But Rigidly

- Geographic weights modify the allocation rule:

$$w_{ij}^G = \exp(-h d_{ij}^G).$$

$$\Pr(z_i = k \mid z_{1:i-1}, \alpha) = \begin{cases} \dfrac{\sum_{j<i} w_{ij}^G \mathbb{I}(z_j = k)}{\alpha + \sum_{j<i} w_{ij}^G}, & \text{existing } k, \\ \dfrac{\alpha}{\alpha + \sum_{j<i} w_{ij}^G}, & \text{new.} \end{cases}$$



Limitation

Geographic proximity alone may not capture structural similarity: nearby regions can have different profiles, while distant regions can have similar profiles.

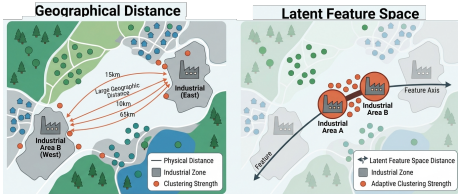
Latent Structural Space

- Let $\boldsymbol{H}_i \in \mathbb{R}^r$ denote observed structural covariates.
- A latent position represents structural similarity:

$$\boldsymbol{\eta}_i = \boldsymbol{B}\boldsymbol{H}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} N_q(0, \tau^2 \boldsymbol{I}_q).$$

Two distinct roles

- $\boldsymbol{x}_i$: covariates in the disease regression.
- $\boldsymbol{H}_i$: structural covariates informing the latent representation used in the clustering prior.



Similar regions → close in latent space.

Adaptive Distance: Combining Geography and Structure

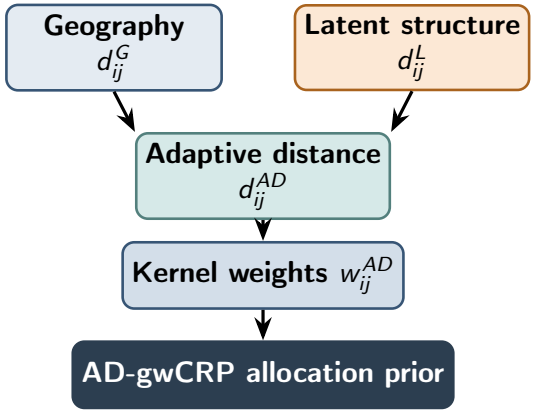
Adaptive distance

$$d_{ij}^{AD} = \gamma \frac{d_{ij}^{G}}{c_G} + (1-\gamma)\frac{d_{ij}^{L}}{c_L}, \qquad 0 \le \gamma \le 1.$$

Kernel weights

$$w_{ij}^{AD} = \exp(-\lambda d_{ij}^{AD}).$$

- d_{ij}^{G}: geographic distance.
- d_{ij}^{L}: latent structural distance.
- λ: controls weight decay with distance.
- γ: balances the two distance components.

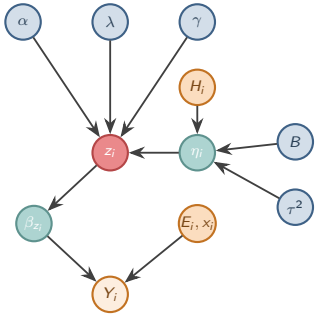


Data-driven The mixing parameter Bayesian adaptation

$\gamma \sim \mathrm{Beta}(a_\gamma, b_\gamma)$ is estimated in MCMC from the data through posterior inference rather than fixed in advance.

Bayesian Inference: Metropolis-within-Gibbs

- Cluster-specific coefficients β_k integrated out via Laplace approximation, making partition updates computationally feasible.
- **MCMC updates:**
  1. Partition z: MH using AD-gwCRP proposals.
  2. Latent positions η_i: random-walk MH.
  3. B, τ^2: Gibbs (Gaussian / Inv-Gamma).
  4. γ: MH on logit scale.
- α (concentration) and λ (kernel bandwidth) are treated as fixed hyperparameters; sensitivity to their values is examined in supplementary analyses.



Application: Swedish Lung Cancer Mortality

Data

- 290 municipalities, 21 counties
- Deaths 2020–2024, age ≥ 25
- 580 municipality–sex observations
- E_i: expected deaths from national crude reference rate (Statistics Sweden / mortality registries).
- Structural Covariates: education, income, physical activity, regional healthcare access, population density

Adaptive distance

$$\widehat{\gamma} = 0.247$$

About 25% geographic and 75% latent structural

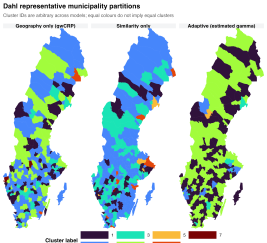
Model comparison

Model	WAIC	LPML	MAE	$\bar{K}$	$\hat{\gamma}$
gwCRP (Geo)	3539	−1771	3.30	5.5	1.00
gwCRP (Sim)	3576	−1778	3.27	6.3	0.00
AD-gwCRP	**3488**	**−1743**	**3.24**	**4.9**	**0.25**

Main findings

AD-gwCRP gives the best WAIC, LPML, and MAE, with the fewest clusters.

Application: Spatial Clusters and Female Relative Risk



Dahl representative partitions

AD-gwCRP yields fewer, more spatially coherent clusters. Labels are not comparable across models.



Fitted female relative risk

$RR_i = \widehat{\mu}_i / E_i$. AD-gwCRP shows a northern high-risk region with $RR \approx 1.5$–1.65.

Conclusion

Combining geographic and latent structural distances improves fit (WAIC, LPML) and gives coherent spatial clusters. In-sample MAE alone does not establish out-of-sample superiority.

Conclusion and Future Directions

- AD-gwCRP combines geographic distance and latent structural similarity in a Bayesian nonparametric clustering framework.
- Estimating γ allows the relative contribution of the two distance components to be learned from the data.
- In the evaluated simulations and application, AD-gwCRP showed favourable results on the reported criteria relative to the selected baselines.
- **Future directions:** out-of-sample validation, spatio-temporal extensions, and scalable Bayesian inference.

Thank You!

Geographic proximity and structural similarity can provide complementary information for Bayesian spatial clustering.

Bayes Bayesian Biostatistics Leiden, NL 2026

UMEÅ UNIVERSITET